HONI SOIT QUI MAL Y PENSE
DIEU ET MON DROIT

AGRICULTURAL RESEARCH INSTITUTE
PUSA

BIOMETRIKA

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS

FOUNDED BY
W. F. R. WELDON, FRANCIS GALTON AND KARL PEARSON

EDITED BY
KARL PEARSON
ASSISTED BY
EGON SHARPE PEARSON

VOLUME XVII
1925

ISSUED BY THE BIOMETRIC LABORATORY
UNIVERSITY COLLEGE, LONDON
AND PRINTED AT THE
UNIVERSITY PRESS, CAMBRIDGE

PRINTED IN GREAT BRITAIN

CONTENTS OF VOLUME XVII

Memoirs

PAGE

I. A Study of Egyptian Craniology from Prehistoric to Roman Times. By G. M. **Morant** 1

Note. The Definition of the Alveolar Point. By **Karl Pearson** . 53

II. On the Moments of the Hypergeometrical Series. By Professor V. **Romanovsky** 57

III. Probability Integrals for a Hypergeometrical Series. By Professor **Burton H. Camp** 61

IV. Determination of $\int_0^\theta \cos^{n+1}\theta \, d\theta$ for Large Values of n, and its Application to the Probability Integral of Symmetrical Frequency Curves. By **John Wishart** 68

V. On the Moments of the Distribution of Squared Standard Deviations for Samples of N drawn from an Indefinitely Large Population. By A. E. R. **Church** 79

VI. The Corpuscle Problem. A Mathematical Study of a Biometric Problem. By S. D. **Wicksell** 84

VII. The Further Theory of Francis Galton's Individual Difference Problem. By J. O. **Irwin** 100

VIII. On the Multiple Correlation of Brothers, being a Note on Mr J. O. Irwin's Memoir, and on my Statement of the Application of Galton's Difference Problem to the Determination of the Degree of Relationship of Brothers, made in August 1902. By **Karl Pearson** 129

IX. Heights and Weights of Patients in Mental Hospitals. By **Rosa Greenwood**, **Cecily M. Thompson**, and **Hilda M. Woods** . 142

X. Further Contributions to the Theory of Small Samples. By **Karl Pearson** 176

XI. James Bernoulli's Theorem. By **Karl Pearson** 201

XII. On the Means, Standard Deviations, Correlations and Frequency Distributions of Functions of Variates. By Professor **Kazutaro Yasukawa** 211

XIII. On a Criterion for the Rejection of Outlying Observations. By J. O. **Irwin** 238

XIV. Notes on an Experimental Test of Errors in Partial Correlation Coefficients, derived from Fourfold and Biserial Total Coefficients. By **Ethel M. Newbold** 251

Note on Miss Newbold's Memoir. By the **Editor** 266

PAGE

XV. The Fifteen Constant Bivariate Frequency Surface. By **KARL PEARSON** 268

XVI. On a Skew Correlation Surface. By **E. C. RHODES** 314

XVII. On the Excess Mortality of Males in the First Year of Life. By Major **GREENWOOD** and **ETHEL M. NEWBOLD** 327

XVIII. Table of the First Twenty Tetrachoric Functions to Seven Decimal Places. By **ALICE LEE** 343

XIX. Modern Problems in Vital Statistics. By Professor **HARALD WESTERGAARD** 355

XX. On the Extreme Individuals and the Range of Samples taken from a Normal Population. By **L. H. C. TIPPETT** 364

XXI. Bayes' Theorem, examined in the Light of Experimental Sampling. By **EGON S. PEARSON** 388

XXII. Mutually Consistent Multiple Regression Surfaces. By **BURTON H. CAMP** 443

Miscellanea

(i) Some further Notes on Cancer and Goitre Distributions. By **PERCY STOCKS** 159

(ii) Recurrence Formulae for the Moments of the Point Binomial. By **RAGNAR FRISCH** 165

(iii) Review: Newsholme's "The Elements of Vital Statistics in their bearing on Social and Public Health Problems." By E. M. E. . 171

(iv) Lebenserwartung und mittleres Alter der Lebenden. Von **E. I. GUMBEL** 173

(v) Table of the Differences of the Powers of Zero. Computed by K. P. and E. M. E. 200

(vi) On First Power Methods of Finding Correlation. By **KARL PEARSON** 459

(vii) Further Remarks on a Previous Paper entitled "Determination of $\int_0^\theta \cos^{n+1}\theta\, d\theta$ for large values of n." By **JOHN WISHART** . . . 469

(viii) Contributions to the Theory of Small Samples drawn from a finite Population. By **J. SPLAWA-NEYMAN** 472

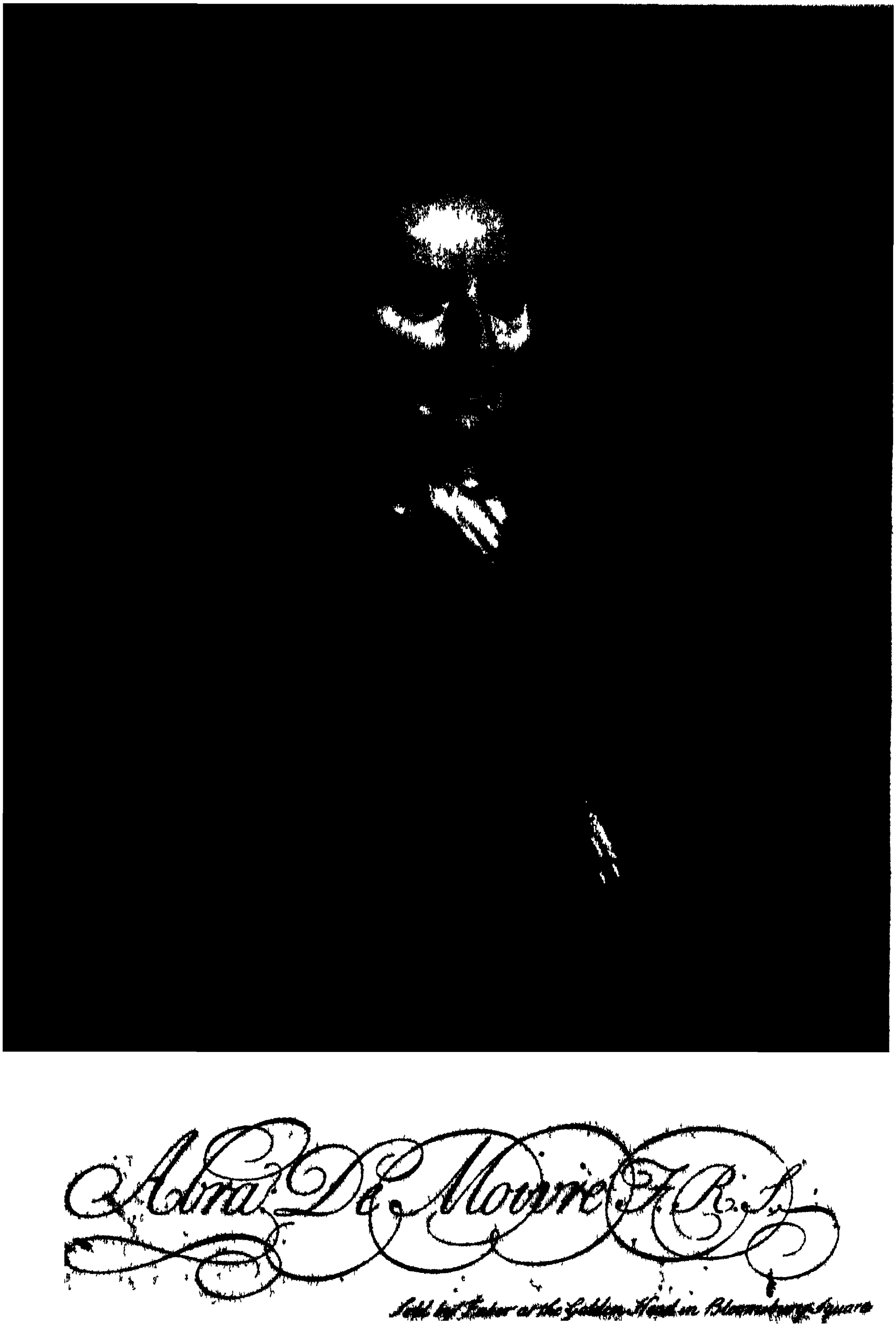
List of Plates, etc.

(α) "Biometrika" Portrait Series, No. I. Abraham de Moivre, Discoverer of the "Normal Curve" . . . . . . *Frontispiece*

"Biometrika" Portrait Series, No. II. Rudolf Martin . . *to face page* 201

(β) Morant's Egyptian Craniology.
Folding Sheet of Measurements of Crania of First Dynasty . . . 52
Six Tissues of Cranial Contours, First Dynasty, in pocket at end of volume.

(γ) Pearson's Fifteen Constant Bivariate Frequency Surface.
Two Tissues of Barometric Height Contours in pocket at end of volume.


Abra: De Moivre F.R.S.
Sold by Faber at the Golden Head in Bloomsbury Square

Abraham De Moivre: Discoverer of the “Normal Curve

Erratum.

Subscribers to *Biometrika* are requested to make the following correction in the paper by Mr Frank Sandon in *Biometrika*, Vol. XVI. upon the "Simplification of the Calculation of Abruptness Coefficients," p. 194, l. 11. In the value for μ_2' a negative sign has been dropped before the term $\frac{2x_p'k_1}{N}$ in the square brackets. A computer recently wasted much time by not observing that the sign of this term was in error.

VOLUME XVII JUNE, 1925 NOS. 1 AND 2

BIOMETRIKA

A STUDY OF EGYPTIAN CRANIOLOGY FROM PREHISTORIC TO ROMAN TIMES.

BY G. M. MORANT, M.Sc., Crewdson-Benington Student in Craniometry.

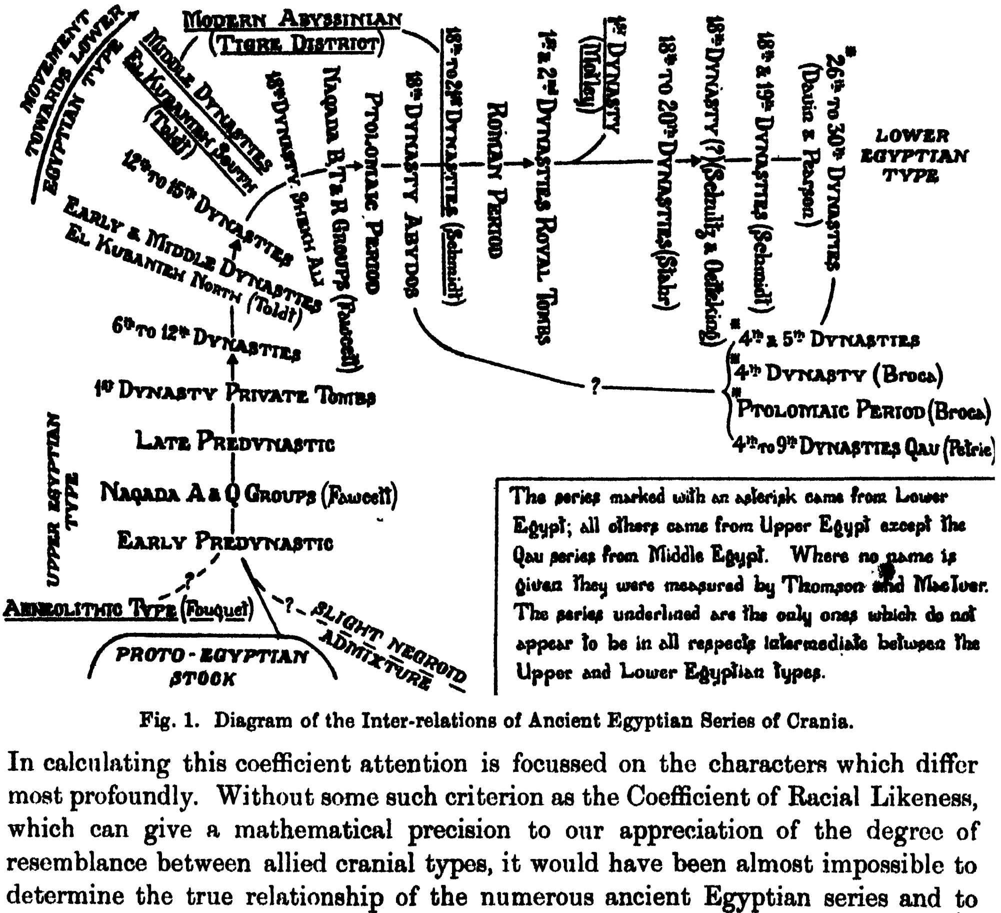
Together with a series of measurements on Crania of the First Dynasty from Royal Tombs at Abydos. By G. H. MOTLEY, B.Sc.

CONTENTS.

PAGE
(1) Introductory 1
(2) Methods of Measurement and Remarks on the Cranial Anomalies of First Dynasty Royal Tombs skulls from Abydos . . . 10
(3) Prehistoric Series of Egyptian Skulls 12
(4) Skulls of the First to the Sixth Dynasties 22
(5) Series from Lower Egypt later than the Fifth Dynasty . . . 29
(6) Upper Egyptian Series later than the Fifth Dynasty measured by Thomson and MacIver 32
(7) Upper Egyptian Series later than the Fifth Dynasty measured by Chantre and Broca 34
(8) Upper Egyptian Series later than the Fifth Dynasty measured by German Craniometricians 35
(9) Type Contours of First Dynasty Skulls from Royal Tombs at Abydos 44
Appendix:
Table of Absolute and Indicial Measurements . . . *Facing* 52

1. *Introductory.* The craniology of the ancient Egyptians has been more intensively studied than that of any other race of man and perhaps no other people has been accredited at different times with such varied origins. Since Morton (1), who supposed them the posterity of Ham, furnished descriptions and measurements of dynastic crania, the craniometric material has been slowly accumulating, although only a few of the series dealt with in more recent years can satisfy all the requirements of biometry. The present paper provides a notable addition to our knowledge in dealing with hitherto undiscussed measurements of two series of crania which are preserved in the Biometric Laboratory. One is of 50 ♂ and ♀ skulls from Royal Tombs of the First Dynasty at Abydos, measured by Mr G. H. Motley, and the other of 900 ♂ skulls from Gizeh belonging to the 26th to 30th Dynasties, measured by Miss A. Davin and other workers in the Laboratory. Both series were excavated by Sir Flinders Petrie. The latter is the only one available which is long enough to give really adequate representations of a statistical population, but there are several others sufficiently long to give quite reliable mean measurements. An attempt has been made to collect from published sources the mean measurements of all series of Egyptian ♂ crania from prehistoric

to Roman times which contain more than 20 specimens*. In estimating the inter-relations of such abundant material the need for the greatest possible refinement in method is at once felt. The differences between the various types are in many cases so small that it is impossible to determine their affinities from a comparison of isolated characters; the true sequences are obscured by errors of random sampling. In such cases the method of the Coefficient of Racial Likeness, which was originally suggested by Professor Karl Pearson and of which extensive use has now been made, becomes an indispensable aid. That criterion takes cognisance at one time of all the principal measurements of the skull and it gives a measure of the relationship between two series compared as a single numerical quantity.


MOVEMENT TOWARDS LOWER EGYPTIAN TYPE
MODERN ABYSSINIAN (TIGRE DISTRICT)
MIDDLE DYNASTIES EL KUBANIEH SOUTH (Toldt)
12th TO 15th DYNASTIES
EARLY & MIDDLE DYNASTIES EL KUBANIEH NORTH (Toldt)
18th DYNASTY: SHEIKH ALI
NAQADA B, T & R GROUPS (Fawcett)
PTOLOMAIC PERIOD
18th DYNASTY ABYDOS
18th TO 20th DYNASTIES (Schmidt)
ROMAN PERIOD
1st & 2nd DYNASTIES ROYAL TOMBS
1st DYNASTY (Morley)
18th TO 20th DYNASTIES (Stahr)
18th DYNASTY (?) (Schultz & Oetteking)
18th & 19th DYNASTIES (Schmidt)
*26th TO 30th DYNASTIES (Davis & Pearson)
LOWER EGYPTIAN TYPE
*4th & 5th DYNASTIES
*4th DYNASTY (Broca)
*PTOLOMAIC PERIOD (Broca)
4th TO 9th DYNASTIES QAU (Petrie)
?
6th TO 12th DYNASTIES
1st DYNASTY PRIVATE TOMBS
LATE PREDYNASTIC
NAQADA A & Q GROUPS (Fawcett)
EARLY PREDYNASTIC
UPPER EGYPTIAN TYPE
AENEOLITHIC TYPE (Fouquet)
?
SLIGHT NEGROID ADMIXTURE
PROTO-EGYPTIAN STOCK

The series marked with an asterisk came from Lower Egypt; all others came from Upper Egypt except the Qau series from Middle Egypt. Where no name is given they were measured by Thomson and MacIver. The series underlined are the only ones which do not appear to be in all respects intermediate between the Upper and Lower Egyptian types.

Fig. 1. Diagram of the Inter-relations of Ancient Egyptian Series of Crania.

In calculating this coefficient attention is focussed on the characters which differ most profoundly. Without some such criterion as the Coefficient of Racial Likeness, which can give a mathematical precision to our appreciation of the degree of resemblance between allied cranial types, it would have been almost impossible to determine the true relationship of the numerous ancient Egyptian series and to reduce them to any uniform plan. The comparison of the Coefficients leads to the arrangement of the types shown in Fig. 1.

---

* It is surprising that so many German craniologists writing in recent years, and with quite a lively sense of the need for full measurements of as many skulls of a homogeneous population as possible, should be content to publish individual measurements without means. I have had to calculate the means of nearly all the material dealt with in this paper with the exception of the series in the Biometric Laboratory.

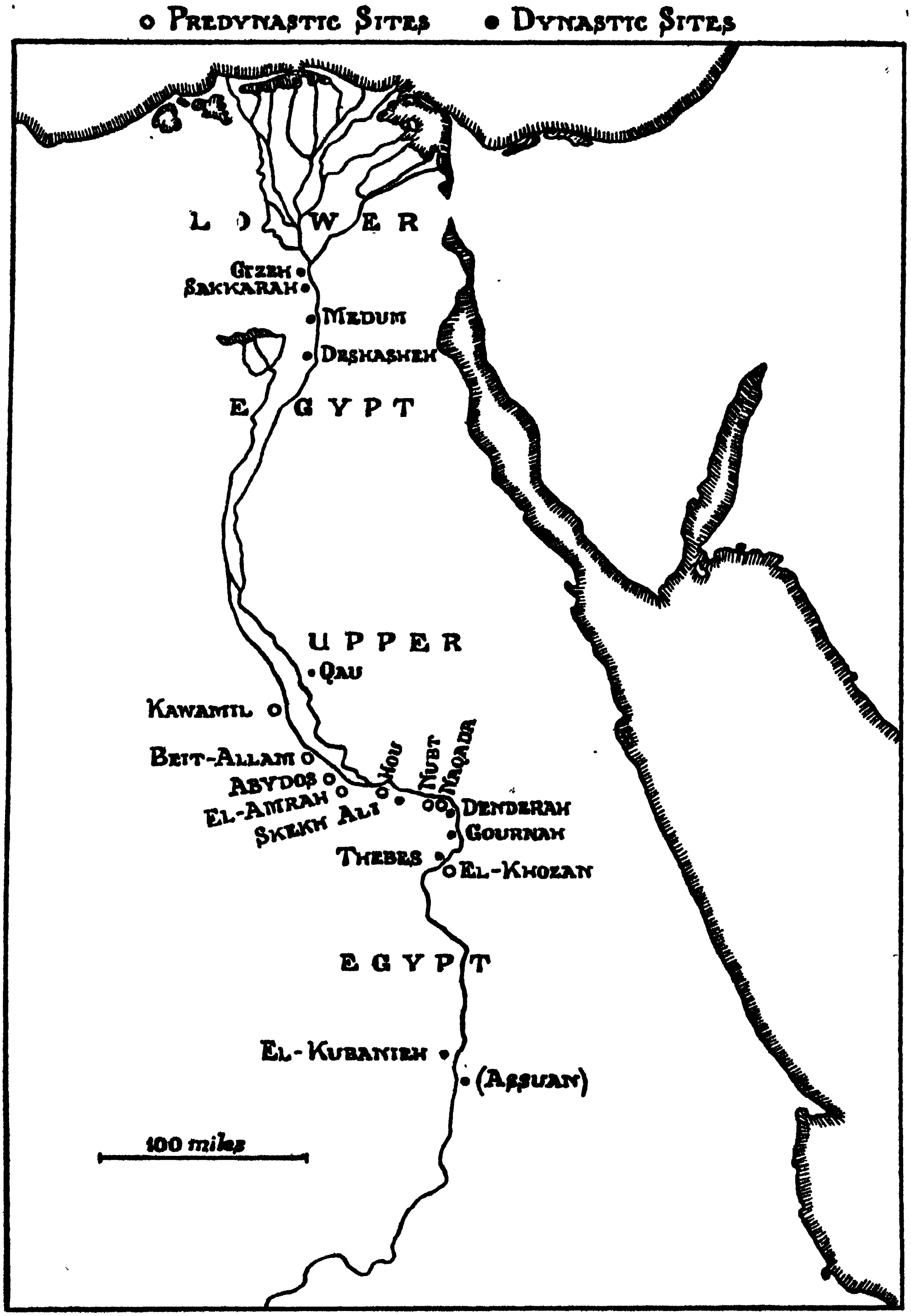


Fig. 2. Map showing sites from which the cranial series dealt with in this paper were obtained.

In his Presidential Address to the Anthropological Section of the British Association in 1923 Professor P. E. Newberry said: "It is remarkable that so little is known about the early history of the Delta. Few excavations have been carried out there, and nothing of Pre-dynastic, or Early Dynastic, times has, so far, been brought to light from the country north of Cairo" ((2), p. 183)*. There are no measurements of Pre-dynastic skulls and very few of Early Dynastic specimens from Lower Egypt, but the long series of 26th to 30th Dynasties skulls from Gizeh makes our knowledge of the physical type prevailing there in later times quite exact. For the district around Thebes and Abydos in Upper Egypt the craniometric material is far more ample than for any other region of the same size in the world.

A detailed comparison of all the ancient Egyptian material leads to the following main conclusions:

(i) In Pre-dynastic times there was a primitive, dolichocephalic race living in Upper Egypt which may have been directly descended from the people whose palaeoliths are found on the high desert floor on either side of the Nile Valley. Though quite distinct from the main population, it bears certain relations to that more advanced type which suggests that the two had a common origin. This so-called Aeneolithic† race may have been assimilated into the main stock of Upper Egypt if its numbers were small, but there is not the slightest suggestion of modification of the skull form of the latter resulting from such an admixture. At any rate the Aeneolithic type had disappeared from the Thebaid before the opening of the historic era.

(ii) In Early Pre-dynastic times there were two distinct races of man living in Egypt; one in the Thebaid and the other, it is supposed, in the Fayum. These may be called the Upper and the Lower Egyptian races. They were as closely related to one another as two adjacent peoples are generally found to be, and there can be no doubt that they diverged from the same branch of the human tree at no very early date. The earliest direct evidence we have of the Lower Egyptian type is in the persons of the royal retainers of the First and Second Dynasties who were buried in Abydos, but it is probable that they were intrusive in Upper Egypt as the type is quite distinct from that of the contemporary native population. The earliest series of Lower Egyptian type found in the Fayum belong to the Fourth Dynasty, but the presumptive evidence in favour of the view that the population there was of almost precisely the same type in much earlier times is very strong. The Upper Egyptian type is known directly from crania found in the Thebaid and dating from Early Pre-dynastic times.

(iii) The Lower Egyptian type seems to have remained unchanged from Early Dynastic to Ptolemaic times except that a relatively small part of the population was modified very slightly, possibly by admixture with some unknown foreign race.

---

* All the sites from which skull series dealt with in this paper were obtained are shown on the accompanying map (Fig. 2).

† In using the term *Aeneolithic*, as in Italian usage, I am adopting a suggestion made by Sir Flinders Petrie.

(iv) The Upper Egyptian type was slowly transformed from the very earliest times in which we have acquaintance with it, and by Late Dynastic times the population of Upper Egypt was of almost pure Lower Egyptian type.

(v) The pure Upper and Lower Egyptian types bear the following relations to one another:

(*a*) All characters measuring lengths in the sagittal plane, except the basio-alveolar length GL (viz. L, LB, S, S_1, S_2, S_3, S_3', fml and G_1'), facial breadth characters ($\mathfrak{Z}B$, O_1', G_2 and NB), $A\angle$ and $B\angle$ of the fundamental triangle, the profile angle, occipital index and palatal index are identically the same for the two types.

(*b*) The facial heights $G'H$ and NH', foraminal breadth (fmb) and upper facial index ($100\,G'H/GB$) of the Lower Egyptian type are slightly greater than those of the Upper, but the height index $100\,H'/L$ is just appreciably smaller.

(*c*) Considerable differences are found between the characters Q', U, J, B' and cranial capacity, all being in this case primarily dependent on differences between the calvarial breadths, and the orbital height and index are greater for the Lower Egyptian type, while, in keeping with the latter difference, the nasal index is less. With its much greater calvarial breadth the Lower Egyptian skull has the smaller basio-bregmatic height and its basio-alveolar length and nasal angle are also smaller than those of the southern type.

(*d*) The greatest differences are between the values of the calvarial breadth, the cephalic index, which is entirely dependent on B as L is constant, and the index $100\,B/H'$, the latter being a peculiarly distinctive character as B and H' bear opposite relations to one another.

We conclude then that the calvaria of the Lower Egyptian skull has a greater breadth, a rather smaller height and the same length as that of the Upper Egyptian type. The facial bones of the two, other than the zygomatic, are equal in breadth but the broader skull has the rather longer face.

These two types represent the extremes of the pure native Egyptian population from Early Pre-dynastic to Ptolemaic times and the relations between them are of the same nature as those which two allied, adjacent and *contemporaneous* races would bear to one another in general. The vast majority of the Dynastic series of crania from Upper Egypt are intermediate between the Upper and Lower Egyptian types. Characters—such as those related to the calvarial length—which are the same for the extremes are constant for all and where the greatest differences are shown the transition from one to the other can be clearly traced. The series represent various stages in the transformation of one type into the other. The types shown in the following table were selected because each represents an adequate number of crania* and shows no signs of intermixture with any race foreign to Egypt. The order in which they are arranged is that deduced from a comparison of the Coefficients of Racial Likeness between them. Every series came from Upper Egypt except that of the 26th to 30th Dynasties which represents the pure Lower Egyptian type.

---

* No mean measurement shown in this table is based on fewer than 36 skulls and all those not marked with an asterisk are based on more than 50.

Series	Measured by	*B*	100*B*/*L*	100*B*/*H'*	*H'*	*J*	100*NB*/*NH'*	*N*∠	*GL*	*L*	*LB*	*A*∠
Early Pre-dynastic	Thomson & MacIver	131·4	71·7	98·1*	134·0*	124·3*	50·3*	67°·0*	98·8*	183·5	102·0*	72°·6*
Naqada *A* & *Q* Groups	Fawcett ...	132·7	71·8	99·2	133·8	125·9*	50·0	65°·5*	96·7*	184·7	101·4	73°·3*
Late Pre-dynastic	Thomson & MacIver	133·4	72·1	100·5	132·7	127·0	51·0	66°·8	98·3	185·1	101·6	72°·4
6th to 12th Dynasties	"	134·3	73·6	99·6	134·8	124·5	50·4	65°·7	96·5	182·5	100·8	72°·8
12th to 15th Dynasties	"	133·8	73·1	100·6	133·0	126·4	48·8	65°·7	96·7	183·0	100·6	72°·0
Early and Middle Dynasties (El Kubanieh South)	Toldt ...	135·9	74·4	102·3	132·9	125·7	50·8	67°·5	98·1	182·7	100·1	71°·0
Ptolemaic Period	Thomson & MacIver	135·3	73·7	102·9	131·5	128·6	49·3	64°·3	95·0	183·5	100·7	73°·3
18th to 21st Dynasties	E. Schmidt	136·6	75·1	103·6	131·9	128·3	50·1	66°·1	96·4	181·9	100·6	73°·0
Roman Period ...	Thomson & MacIver	136·4	75·2	104·4	130·7	127·5*	49·4*	64°·3	94·2	181·4	99·9	73°·2
18th to 20th Dynasties	Stahr ...	137·3	74·8	102·7	133·7	128·6	50·4	64°·5	96·1	183·5	101·9	74°·0
26th to 30th Dynasties	Davin and Pearson	138·9	75·1	104·9	132·4	128·7	47·3	63°·9	95·0	185·3	101·6	74°·1

*** No mean measurement shown in this table is based on fewer than 36 skulls and all those not marked with an asterisk are based on more than 50.**

The differences between adjacent types in the above table are so small that it is surprising to find that the sequences of characters are as regular as they appear to be. If the numbers on which the mean measurements were based had been smaller the orders would very probably have been hopelessly obscured for the characters showing the smaller differences between the extremes. All the characters show significant differences except *L*, *LB* and *A* ∠ on the extreme right of the table and the constancy of those three is very striking.

(vi) The question how exactly the population of Upper Egypt was transformed from its pristine type to that of the Lower Egyptian race is one of considerable importance from both an evolutionary and an historical point of view. To answer it with any degree of finality we should require the evidence not of 5000 but of 50,000 skulls and our present conclusions can be no more than tentative. The available series are not scattered randomly in time: there are very few for Pre-dynastic and Early Dynastic times, the later dynasties are well represented but the evidence relating to Ptolemaic and Roman times is meagre. The 18th to 20th Dynasties are far more adequately represented than any other period of the same length*. If the Pre-dynastic period was as long as the Dynastic, as is generally

---

*** That is by the number of cranial series, but if the number of skulls is considered the 26th to 30th Dynasties are far better represented than any other period as the long E series belongs to that time.**

supposed, then the change in the Upper Egyptian type prior to the First Dynasty was exceedingly slow, but quite uniform as far as we can tell. Such a conclusion is quite in keeping with the archaeological evidence. In the early dynasties the movement became more rapid and by the 15th the skull form of the natives of Upper Egypt was approximately mid-way between those of the Upper and Lower Egyptian types. Either during or immediately preceding the 18th Dynasty there seems to have been a considerable acceleration of the movement. Four series belonging to the 18th to 20th Dynasties—two from Thebes and two from Abydos—are of almost pure Lower Egyptian type, but like the Royal Tombs series of the First and Second Dynasties, to which they bear the closest possible resemblance, they still incline slightly towards the Upper Egyptian type. It is possible that those five series indicate sudden intrusions into Upper Egypt which were independent of the slower movement*. There are two other 18th and 19th Dynasties series which represent the slower and more normal transformation which pursued its course steadily down to Ptolemaic and Roman times. The rate of change after the First Dynasty appears to have been on the whole remarkably uniform and there is no suggestion of a retrograde movement—that is to say a reversion to the Upper Egyptian type—at any time.

(vii) As examples of types which were entirely unaffected for several thousand years by any influences foreign to their country, the ancient Egyptians may well be unparalleled in the history of the world. All the series which do not appear to be of absolutely pure Egyptian type are underlined in Fig. 1 and, with the exception of Fouquet's Aeneolithic type, the divergences from the main stock are exceedingly small†. The most aberrant series is Motley's First Dynasty from the Royal Tombs of Abydos. It is most similar to the Lower Egyptian type, though inclining slightly towards the Upper Egyptian—the royal retainers being supposed intrusive in the Thebaid—but it also shows signs of admixture with an alien race which was certainly not Negroid The calvariae of these First Dynasty skulls are not in any way peculiar, but the facial heights, and to a less marked extent the facial breadths, are characteristically large. No other series dealt with in this paper has shown any tendency to diverge in the same direction. Three series from Lower Egypt are of pure Lower Egyptian type for all characters except the basio-bregmatic height which has values which are quite significantly greater than that of the 26th to 30th Dynasties type‡. An 18th Dynasty series from Abydos, which in all other ways is intermediate between the Upper and Lower Egyptian types, is also characterised by a large basio-bregmatic height. Judging from means reduced by Lee from Schmidt's individual measurements (see (6), p. 426), the modern Copts have diverged greatly from the type of the dynastic Egyptians and in no way more significantly than in their reduced length ($L = 177{\cdot}0$) and greater cephalic index ($100B/L = 77{\cdot}3$). But the modern natives of Northern Abyssinia measured by Sergio Sergi (26) are surprisingly similar to the type which predominated in Upper Egypt during the middle dynasties. The relationship between one of the latter and the Abyssinians is decidedly closer than that between two series—the Whitechapel and Moorfields—of seventeenth century Londoners. This is a striking example of the persistence of a type with only slight modification for a period of at least 3000 years. There are two middle dynastic series which diverge slightly from the pure native population contemporaneous with them in the ways which were more accentuated in the later Abyssinians.

---

* It is quite probable that the 18th Dynasty series which closely resemble the pure Lower Egyptian type were representatives of the middle and upper classes of the population, while the lower social orders were closer to the pure Upper Egyptian type. Thomson and MacIver have provided measurements of two 18th Dynasty series. One from Shekh Ali was found in a pit and it undoubtedly represented an inferior class; the other, from Abydos, was obtained in a middle class cemetery. The Shekh Ali series is distinctly closer to the earlier type of Upper Egypt than the other. (See (13), p. 29.)

† At different times during the Dynastic era there were undoubtedly quite large numbers of foreign slaves in the country. The only skull measurements of such people available seem to be those of pygmy crania of the Third Dynasty (8).

‡ Sir Flinders Petrie's field measurements of Fourth to Ninth Dynasties skulls from Qau indicate a precisely similar divergence (see p. 29, ftn. † below).

(viii) It is very generally supposed that the population of ancient Egypt was sensibly affected at various times by the infusion of Negro blood. But in the series of which we have the mean measurements it is not possible to detect the slightest effect of any such admixture that can have taken place after early Pre-dynastic times. Apart from isolated negroid skulls which are said to have been excluded from several of the series, the populations appear to be quite homogeneous and we have no reason to suppose that the mean type was affected in the slightest by admixture with any race foreign to Egypt. But it is quite possible that the race we have called the pure Upper Egyptian had originally some Negroid elements in its make-up. There are very few measured series of Negro skulls which are sufficiently long to give reliable determinations of type. A comparison of the more reliable material shows that there is a considerable variation in calvarial measurements among Negro races: all the Egyptian series fall within the extremes of length, breadth and cephalic index shown by the Negro types. But certain other measurements—chiefly facial—apparently distinguish all Negro from all Egyptian types. In the table on p. 9 the Upper Egyptian type is represented by the Early Pre-Dynastic and Naqada series and the Lower Egyptian by the 26th to 30th Dynastic series. The Negro races were selected because more or less adequate skull measurements have been furnished for them but for no other reason. All the measurements are of ♂ skulls.

The characters $NH(R)$, NB, $NB/NH(R)$, $P\angle$, $N\angle$ and $B\angle$ clearly distinguish the Egyptian from all Negroid types: they seem to be very constant for the latter. At the same time we find that for each one of those measurements the Upper Egyptian types are nearer to the Negro values than are the Lower Egyptian, and the Early Pre-dynastic is very slightly nearer to the Negroes than is the Naqada type. Much the same relations are found for the basio-alveolar length (GL), although the low value for the Congo series is discordant, and the indices $100H'/L$ of the Upper Egyptian types are also between the Lower Egyptian and Negro values. With the facial height ($G'H$) and index ($100G'H/GB$) the case is rather different. For those two characters the Bantu Negroes are almost identically the same as the ancient Egyptians, but the smaller values of the non-Bantu races differentiate that

		Number of Skulls	*NH, R*	*NB*	100 *NB/NH, R*	*P* ∠	*N* ∠	*B* ∠	*GL*	*G'H*	100 *G'H/GB*	100 *H'/L*	S_3'	S_3	*Oc. I.*
Ancient Egyptians	26th to 30th Dynasties	850	51·7	24·4	47·3	85°·8	63°·9	41°·9	95·0	70·4	74·1	71·5	94·7	115·2	61·5
	Naqada *A* and *Q* Series	70	50·2	25·0	49·8	84°·6	65°·5	41°·2	96·7	69·7	72·7	72·4	96·5	116·9	60·2
	Early Pre-dynastic	40	50·1	25·2	50·3	—	67°·0	40°·4	98·8	69·2	—	73·3	—	—	—
Bantu Negroes	Zulu* ...	20	47·2	27·3	58·1	—	70°·1	39°·8	101·9	69·5	72·1	75·0	—	111·7	—
	Angoni* ...	25	48·5	28·0	58·0	—	70°·5	39°·9	103·5	70·3	74·1	74·3	—	108·9	—
	Kaffir* ...	35	48·6	27·3	56·3	—	70°·7	38°·6	105·1	69·1	72·2	72·1	—	119·3	—
Non-Bantu Negroes	Congo* ...	50	47·2	26·0	55·2	82°·8	69°·3	37°·9	96·5	63·4	66·9	75·4	95·0	107·7	68·8
	Gaboon, 1864 Series*	50	48·2	26·6	55·3	—	70°·5	38°·6	100·5	66·4	69·7	75·5	95·9	109·1	68·3
	Northern Negro* ...	30	49·5	27·3	55·4	82°·0	72°·4	38°·0	105·3	68·0	70·4	74·2	—	—	—
	Cameroons† .	80	49·1	27·5	56·0	80°·5	72°·0	39°·1	102·6	67·9	68·1	75·0	97·6	112·4	65·6
	East Coast (Dschagga)‡	12	44·3	27·8	62·2	—	73°·0	38°·6	105·0	68·2	70·5	73·2	—	—	—

* The series marked with an * were taken from (7), pp. 298, 299. The Zulu, Angoni and Kaffir means were reduced from individual measurements provided by Shrubsall. The Congo and Gaboon series were measured by Crewdson-Benington and the Northern Negroes were taken from various parts of the German Anthropological Catalogue (19). Of the last series it was said it can "only be of service when we are comparing negroes in general with other races": see (7), p. 295.

† The means of these Cameroons' skulls were reduced from individual measurements provided by Drontschilow (25). In calculating them I omitted 7 pathological and 3 young skulls.

‡ The individual measurements of these Dschagga skulls are provided by Widenmann (24).

group from the others although they are more similar to the Upper than to the Lower Egyptian values. The flattened occipital bone of the non-Bantu Negro skull is a most characteristic feature but we cannot say that in that respect one Egyptian race is more negroid than the other. From a comparison of facial measurements alone we can at least say that there is a distinct suggestion of negroid affinities in what has been called the Upper Egyptian race. The calvarial measurements do not support that view, and we cannot gauge the significance of the observed relationships while any knowledge relating to the skull form of a type which has resulted from the crossing of two races as markedly different as a pure Negro and a pure Egyptian is quite lacking.

In dealing with the long-bones of the Naqada series Warren (9) found that there was a suggestion of negroid affinities[*].

(ix) A study of all the series of early Egyptian crania that have as yet been measured shows that they are inter-related in a simple way. It is most probable that the observed differences between the types are to be explained as being due to some slow moving evolutionary process—whether through selection or directly through environment—or to a slow blood mingling of different races. On the latter hypothesis it would only be necessary to assume the existence of two original races, an Upper and a Lower Egyptian, which were distinct in Early Pre-dynastic times. The gradual approach of the Upper Egyptian population to the Lower Egyptian type—a movement, it should be noted, which was not completed as late as Roman times—would then be due to admixture in the Thebaid, while the original northern type was apparently persisting in the Fayum without modification. On the other hand, an evolutionary hypothesis would give an explanation of the facts which would be equally reasonable. Additional evidence, however, such as that which would result from a study of the variabilities of the series dealt with in this paper or the discovery of an adequately long Pre-dynastic cranial series in Lower Egypt, might tell strongly in favour of the one hypothesis or the other. It may be that the observed changes were the result of both racial admixture and an evolutionary process.

2. *Methods of Measurement and Remarks on the Cranial Anomalies of First Dynasty Royal Tombs skulls from Abydos.* The individual measurements taken by Mr G. H. Motley and "remarks" on the more outstanding cranial anomalies are contained in the Appendix to this paper. The methods of measurement used in dealing with the First Dynasty and the 26th to 30th Dynasties series from Gizeh, which was measured by Miss A. Davin and others, are precisely those which are ordinarily taken in the Biometric Laboratory[†]. The majority are in accordance

---

[*] **Warren says: "On the whole, the proportions of the limb-bones to one another may be said to have approached those of the Negro, while the sacral and scapular indices were almost identical with those of Europeans." ((9), p. 191.)**

[†] **Full descriptions of the measurements are given in *Biometrika*, Vol. xiv. pp. 196—200. Briefly:**

F=Ophryo-occipital length, L'=glabellar projective length, L=glabella-occipital length, B=maximum parietal breadth, B'=least forehead breadth, H'=basio-bregmatic height, H=Frankfurt vertical height, OH=Frankfurt auricular height, LB=basion to nasion, Q=transverse arc through "apex" terminating at ear-rods of craniophor, Q'=similar to Q terminating at auricular points, S=sagittal arc

with those of the *Frankfurter Verständigung.* Two additional measurements are given for the First Dynasty and for the re-measured Naqada skulls (see notes to Table I). One is the measurement of nasal height (NH') from the nasion to the base of the anterior nasal spine which is used by Broca, Flower and the majority of French and English craniometricians. NH' is slightly less than the Frankfurt measurements NH, R and NH, L. The other (EOW) may be called the external ocular distance: it is taken between the points where the borders of the orbital ridges right and left meet the fronto-malar sutures (see pp. 22 and 23 below). EOW is the same as Martin's *innere orbitale Gesichtsbreite** : that and the *äussere orbitale Gesichtsbreite* are occasionally taken by German craniometricians.

Remarks on the individual crania of the First Dynasty series are given in the Appendix. Unless otherwise stated each skull is complete with a mandible and of adult age. The following characteristics and anomalies were looked for.

Signs of age. Condition of sutures, development and loss of teeth and the falling in or thinning of the calvaria.

Condition of teeth. The teeth of this series were remarkably well preserved though all were worn. Many of the skulls that were ageing had not lost a single tooth. Hardly any cases of irregular formations round the palate were found and there was only one carious tooth.

Palate. Of the 48 skulls only one has bony ridges across the palatine grooves leading from the pterygo-palatine canals. This is an unusually low percentage but it is quite possible that ridges originally possessed by other specimens had been broken.

Pre-condyles. No pre-condyles were found at the basion or opisthion but there were several cases of projecting spines of bone in other regions.

Tympanic perforation. All cases were recorded.

Base of the pyriform aperture. This feature is indicated by the letters P.B. The floor is described as "flat"—i.e. parallel to the Frankfurt horizontal plane—"sloping upwards," meaning sloping upwards towards the pyriform border, or

---

nasion to opisthion, S_1 = arc nasion to bregma, S_2 = arc bregma to lambda, S_3 = arc lambda to opisthion, S_3' = chord lambda to opisthion, U = horizontal circumference, PH = alveolar point to tip of anterior nasal spine, $G'H$ = nasion to alveolar point, GB = breadth between lowest points on zygomatic-maxillary sutures, J = zygomatic breadth, NH, R and L = nasion to lowest edge of pyriform aperture R and L, NB = nasal breadth, DS = dacryal subtense, DC = dacryal chord, DA = dacryal arc, SS = simotic subtense, SC = simotic chord, O_1, R and L = breadths of orbits using Fawcett's curvature method, O_1' = orbital breadth from dacryon, O_2, R and L = heights of orbits, G_1 = palate length from point of *spina nasalis posterior* to an imaginary line tangential to the inner rims of the alveoli of the middle incisors, G_1' = similar to G, from base of spine, G_2 = palate breadth between inner alveolar walls at second molars, GL = basion to alveolar point, fml = length of *foramen magnum*, fmb = breadth of *foramen magnum*, $P\angle$ = Frankfurt profile angle, $Oc. I.$ = occipital index, $N\angle$, $A\angle$ and $B\angle$ found from lengths LB, $G'H$ and GL with aid of Pearson's Trigonometer, $\theta_1 = 180° - P\angle - N\angle$, $\theta_2 = P\angle - A\angle$, EH and EB = height and breadth of palate using Pearson's Uraniscometer. The capacities were found by the method described by Macdonell (*Biometrika*, Vol. III. pp. 203—206).

* Rudolf Martin, *Lehrbuch der Anthropologie*, p. 552, Ed. 1914.

"sloping downwards." The edge is said to be sharp, flat or rounded. Excluding the Negro skull, the following frequencies are found for the combined ♂ and ♀ First Dynasty Egyptians.

		Edge of Pyriform Aperture			Totals
		Sharp	Blunt	Rounded	
Floor of Pyriform Aperture	Sloping upwards ...	12	4	0	16
	Flat...	9	7	2	18
	Sloping downwards.	5	2	2	9
	Totals ...	26	13	4	43

Asymmetry. *SL* or *SR* denotes whether the Sylvian depression is greater on the right or left side and in the same way *JR* and *JL* refer to the relative sizes of the jugular foramina. The following frequencies are found and the balance in favour of the right side is similar to that usually found in racial series.

	Sylvian Depression	Jugular Foramen
Greater on right side	14	20
Equal	28	10
Greater on left side...	3	10

There were few cases of calvarial or facial asymmetry.

Malar bones. No horizontal sutures across the malar bones were found.

Metopism. There were no cases of the persistent frontal suture.

Ossicles in the sutures. These, as ordinarily, were confined almost entirely to the lambdoid suture. There was one ossicle of the λ.

Conformation at the pterion. All cases of epipteric bones and processes of the temporal articulating with the frontal were recorded.

Interparietal bones. There was one case of the simple interparietal aud one of the tri-partite interparietal with the *os pentagonale* fused into the supra-occipital.

Other features. Cases of flattening of the obelion are recorded. In a number of the skulls a slight post-coronal depression was observed, but it was not very marked in any case. These skulls seem to be more symmetrical and to exhibit fewer anomalies than most racial series.

3. *Prehistoric Series of Egyptian Skulls.* The earliest series of prehistoric Egyptian crania of any length, and of which measurements have been provided, belong to a period in which metal was used. Palaeolithic implements have been found in abundance on the terraces of the Nile, but it may be doubted whether the country ever passed through a true Neolithic phase. The opinion of Sir Flinders Petrie ((4), p. 28) is that as soon as the sediment of the river had been deposited to a sufficient depth to make agriculture possible, an alien people—supposed Libyan—who were already acquainted with the use of copper, invaded the country and settled along the banks of the Nile. The autochthonous population at that time appears to have been one of palaeolithic hunters and the

invaders did not exterminate either the people themselves or their primitive culture. It is unfortunately not yet possible to determine with any great accuracy the chronological sequence of the various Pre-dynastic series of crania that are available. In the case of the prevailing type an order can be given by using the method of Petrie's sequence dates (see (4), pp. 4—12) which are determined chiefly from a comparative study of the pottery; but it is apparently not possible to establish with certainty the relative ages of the burials of the primitive peoples who maintained their lower stage of culture though living in close proximity to the invaders. All the skulls that can be dealt with came from a comparatively small area of Upper Egypt round Thebes and Abydos. There was another centre of civilisation in Pre-dynastic times in the Fayum*, but no measurements of early skulls from that region are available and there are none of the contemporaneous populations of Middle Egypt.

The affinities of the series of skulls may be estimated by comparing the Coefficients of Racial Likeness† between their mean measurements which are shown in Table I (p. 14). In the majority of cases the numbers of skulls on which the means are based are fairly adequate, but there are few series for which more than half of the 31 characters which should ideally be used have been provided. It is important to remember that a generalised measure of relationship which does not take into account several important regions of the skulls may be misleading and particular caution has to be taken when comparing such measures with those calculated for all the 31 characters.

In the appendix to the first volume of J. de Morgan's researches into the origin of Egyptian civilisation (10) Dr Fouquet gave individual measurements of 11 skulls that had been disinterred at El-Amrah near Abydos‡. From the contracted position of the bodies and the nature of the grave furniture, which comprised abundant pottery but very little metal, it was reasonable to infer that the burials were of Pre-dynastic date though from the description of the artefacts that has been given it is not possible to assign them to any particular period of that era. The mean measurements of the nine specimens are given in the second column of our Table I. The series in columns 3, 4 and 5 of that table are the longest provided by Fouquet§ in the second volume of de Morgan's researches (11) and they were thought to be contemporary with the El-Amrah skulls. It is clear that the four series represent very similar if not identical types. The numbers of skulls in each series are small but the method of the Coefficient of Racial Likeness may profitably be applied. The Coefficients in Table II (p. 16) were calculated for 19 characters of which six were indices. The numbers in brackets are the mean numbers of skulls available for those 19 characters in the case of each series.

---

* A map showing all the principal Pre-dynastic stations in Egypt is given by J. de Morgan ((10), p. 68).

† See *Biometrika*, Vol. XVI. pp. 11—14 for a definition of the Coefficient of Racial Likeness and an account of the precautions that have to be taken when applying it in practice.

‡ All sites referred to in this paper are shown on the accompanying map (Fig. 2, p. 8).

§ Two series from Naqada North and Guebel-Silsileh ((11), pp. 326 and 386 respectively) contain fewer than five skulls each and their means are not sufficiently like those of the longer series to make it safe to pool them.

TABLE I.

Mean Measurements of Pre-dynastic Egyptian Series. Males.

Pre-dynastic Date	Early?	Early?	Early?	Early?	Early?	Middle?	?	Early	Late	Late?‖
Locality... ...	El-Amrah	Kawamil	Beit-Allam	Naqada South	Kawamil, Beit-Allam and Naqada, South	Naqada (*A* and *Q* Series)	Naqada (*B*, *T*, and *R* Series)	Abydos, El-Amrah and Hou	El-Amrah and Hou	El-Khozan
Measured by	Fouquet	Fouquet	Fouquet	Fouquet	Fouquet	Fawcett and Morant	Fawcett and Morant	Thomson and MacIver	Thomson and MacIver	Chantre
F	—	—	—	—	—	184·1 (71)	182·9 (21)	182·6 (53)	183·5 (127)	—
L	189·2 (9)	187·9 (19)	189·3 (13)	188·2 (28)*	188·3 (60)	184·7 (101)	183·9 (24)	183·5 (54)	185·1 (127)	182 (24)
B	137·6 (9)	138·9 (18)	134·2 (13)	137·3 (28)	137·1 (59)	132·7 (88)	136·1 (22)	131·4 (53)	133·4 (125)	133 (24)
H'	137·0 (8)	137·1 (16)	138·5 (7)	136·9 (27)	137·2 (50)	133·8 (70)	132·2 (18)	134·0 (40)	132·7 (107)	130 (24)
B'	93·6 (9)	96·3 (19)	98·2 (13)	98·3 (28)	97·6 (60)	92·6 (101)	91·8 (21)	—	—	92 (24)
S_1	132·1 (9)	129·2 (19)	128·7 (13)	125·9 (28)	127·6 (60)	127·7 (100)	127·4 (22)	—	—	—
S_2	136·2 (9)	132·3 (19)	135·7 (13)	134·1 (28)	133·9 (60)	130·2 (95)	129·7 (25)	—	—	—
U	512·6 (9)	520·8 (18)	516·1 (11)	515·2 (28)	517·1 (57)	510·4 (90)	512·3 (18)	—	—	—
LB	103·4 (9)	106·1 (12)	105·3 (6)	105·2 (27)	105·5 (45)	101·4 (69)	100·1 (18)	102·0 (36)	101·6 (99)	—
$G'H$	—	—	—	—	—	69·7 (52)	69·6 (18)	69·2 (39)	70·0 (97)	—
GL	—	—	—	—	—	96·7 (46)	94·3 (16)	98·8 (36)	98·3 (97)	—
J	135·6 (5)	130·0 (1)	132·0 (4)	132·3 (18)	132·1 (23)	125·9 (37)	126·2 (15)	124·3 (34)	127·0 (87)	121 (24)
NH'	53·3 (6)	52·3 (12)	51·2 (8)	49·7 (23)	50·7 (43)	50·0 (54)	50·3 (19)	50·1 (37)	50·1 (102)	48 (24)
NB	24·3 (7)	25·7 (12)	24·4 (7)	25·4 (23)	25·3 (42)	25·0 (64)	25·1 (18)	25·2 (37)	25·5 (102)	25 (24)
DC	20·7 (7)	21·1 (12)	20·5 (7)	22·2 (22)	21·6 (41)	—	—	—	—	24 (24)
O_2	34·2 (7)	33·2 (13)	33·9 (8)	32·5 (26)	32·9 (47)	32·2 (57)	32·1 (19)	—	—	32 (24)
O_1'	38·5 (7)	39·2 (13)	39·3 (8)	39·9 (26)	39·6 (47)	39·1 (57)	39·1 (19)	—	—	38 (24)
fml	34·9 (8)	37·1 (14)‡	39·0 (7)	36·0 (27)	36·8 (48)	35·6 (70)	35·5 (19)	—	—	34 (24)
fmb	29·8 (8)	30·1 (14)	29·9 (8)	29·8 (26)	29·9 (48)	29·5 (70)	29·7 (18)	—	—	28 (24)
100 B/L	72·8 (9)	73·1 (18)	70·6 (13)	72·7 (28)	{72·8 (59)}	{71·8 (88)}	{74·0 (22)}	{71·7 (53)}	{72·1 (125)}	73·4 (23)
100 H'/L	{72·4 (8)}	72·3 (16)	73·9 (7)	72·8 (27)	{72·9 (50)}	{72·4 (70)}	{71·9 (18)}	{73·3 (40)}	{71·8 (105)}	{71·4 (24)}
100 B/H'	{100·3 (8)}	{101·3 (16)}	{96·9 (7)}	{100·3 (27)}	{99·9 (50)}	{99·2 (70)}	{103·0 (18)}	{98·1 (40)}	{100·5 (107)}	{102·3 (24)}
100 $(B-H')/L$	{0·3 (8)}	{1·0 (16)}	{−2·3 (7)}	{0·2 (27)}	{−0·1 (50)}	{−0·6 (70)}	{2·1 (18)}	{−1·6 (40)}	{0·3 (105)}	{1·6 (24)}
100 NB/NH'	{45·6 (6)}	49·0 (12)	48·5 (7)	51·7 (23)	{49·9 (42)}	{50·0 (54)}	{49·9 (18)}	{50·3 (37)}	{51·0 (101)}	{52·1 (24)}
100 O_2/O_1'	{88·8 (7)}	85·4 (13)§	85·8 (8)	82·1 (26)	{83·1 (47)}	{82·4 (57)}	{82·1 (19)}	—	—	{84·2 (24)}
100 fmb/fml	{85·4 (8)}	82·3 (13)	77·6 (5)	82·7 (26)†	{81·3 (48)}	{82·9 (70)}	{83·7 (18)}	—	—	{82·4 (24)}
$N\angle$	—	—	—	—	—	{65°·5 (46)}	{64°·3 (16)}	{67°·0 (36)}	{66°·8 (97)}	—
$A\angle$	—	—	—	—	—	{73°·3 (46)}	{73°·9 (16)}	{72°·6 (36)}	{72°·4 (97)}	—
$B\angle$	—	—	—	—	—	{41°·2 (46)}	{41°·8 (16)}	{40°·0 (36)}	{40°·8 (97)}	—
Capacity	[1473·3 (8)]	[1476·4 (16)]	[1460·5 (7)]	[1465·6 (27)]	[1466·8 (50)]	1388·6 (61)	1382·8 (18)	[1384·0 (40)]	[1396·2 (107)]	[1361·6 (24)]

Notes and footnotes to Table I.

The indices and angles in curled brackets were determined from the mean measurements of the component lengths instead of from the individual measurements. The capacities in square brackets were reconstructed by the formula given by Lee and Pearson (*Phil. Trans.* Series A, Vol. 196, p. 247): $C = \cdot 000266 \times L \times B \times H + 524\cdot 6$ c.cs. H' was used in place of the slightly larger H, so the calculated capacities are likely to be rather smaller than the true ones.

Through the kindness of Professor Karl Pearson, I was able to examine the Naqada crania dealt with by Fawcett (6) and to take several measurements that had not been previously found. When first brought to England the skulls were in a very fragile state and it is no longer possible to take any measurements of several [illegible] that were included in Fawcett's tables of individual measurements. Excluding those skulls, I found a close agreement between Fawcett's measurements and my own for all characters except the following (see pp. 18—19 below):

	B	B'	LB	$G'H$	GL	NH, R	G_1	O_1R
Fawcett	184·5 (122)	91·2 (130)	99·9 (100)	67·3 (80)	95·0 (68)	49·1 (84)	55·0 (74)	42·5 (82)
Morant	183·4 (110)	92·5 (122)	101·1 (87)	69·7 (70)	96·1 (62)	50·3 (73)	52·0 (66)	41·6 (76)

The disparity between the numbers of skulls on which the means are based cannot account for the whole of the differences between them and we are bound to accept the conclusion that the measurements were taken in slightly different ways. In this paper I have used my own measurements, and, besides the above characters, the angles of the fundamental triangle, the indices involving B, $100\,G'H/GB$, $100\,NB/NH, R$, $100\,O_2/O_1R$ and $100\,G_2/G_1$ are altered from their original values. In addition to the means in Table I the following are now available for the two Naqada groups (see pp. 18, 19 below):

	L'	H	OH	S	S_3	S_3'
A and Q Series ...	185·4 (87)	136·4 (71)	115·6 (98)	372·6 (90)	116·9 (85)	96·5 (84)
B, T and R Series	182·8 (20)	134·7 (23)	113·8 (26)	372·3 (20)	116·9 (23)	97·0 (23)
	Q'	$G'H$	GB	NH, R	PH	KOW
A and Q Series ...	302·9 (87)	69·7 (52)	95·9 (60)	50·2 (54)	20·1 (56)	96·1 (56)
B, T and R Series	307·9 (18)	69·6 (18)	95·1 (18)	50·5 (19)	19·9 (17)	94·4 (18)
	G_1	G_1'	G_2	O_1R	O_2R	$100\,O_2/O_1R$
A and Q Series ...	52·5 (49)	48·0 (49)	40·7 (48)	41·6 (57)	32·2 (57)	{77·4 (46)}
B, T and R Series	50·7 (17)	46·1 (17)	40·3 (19)	41·7 (19)	32·1 (19)	{77·0 (19)}
	$100\,NB/NH, R$	$100\,G_2/G_1$	$100\,G_2/G_1'$	$100\,G'H/GB$	$100\,B/L'$	$100\,H/L'$
A and Q Series ...	{49·8 (54)}	{77·5 (48)}	{84·8 (48)}	{72·7 (52)}	{71·6 (87)}	{73·6 (71)}
B, T and R Series	{49·7 (18)}	{79·5 (17)}	{87·4 (17)}	{73·2 (18)}	{74·5 (20)}	{73·7 (20)}
	$100\,H/L$	$100\,B/H$	$P\angle$	θ_1	θ_2	$Oc. I.$
A and Q Series ...	{73·6 (71)}	{97·3 (71)}	84°·6 (50)	{29°·9 (46)}	{11°·3 (46)}	60·2 (84)
B, T and R Series	{73·2 (23)}	{101·0 (22)}	86°·1 (16)	{29°·6 (16)}	{12°·2 (16)}	60·5 (23)

* Wrongly given by Fouquet as 184·6.

† Wrongly given by Fouquet as 88·1.

‡ „ „ „ 87·7.

§ „ „ „ 86·8.

‖ The means in this column are given by Chantre (15) in whole mms. only and it is probable that they were not calculated further, so they are not necessarily correct to the nearest mm. The mean indices were calculated from these curtailed lengths and they may easily differ from the true means by as much as one unit. There were 24 ♂ skulls in the series but some lengths, and in particular H' and J, could almost certainly not be taken on all of them. Chantre does not say how many skulls each mean was based on. The material is presented in such a thoroughly unsatisfactory way that it is practically worthless.

TABLE II.

Coefficients of Racial Likeness between Fouquet's Pre-dynastic Series.*

		El-Amrah (7·7)	Kawamil (14·2)	Beit-Allam (8·3)	Naqada South (25·9)
El-Amrah (7·7) ...	19 Characters .. 6 Indices ...	— —	0·34 ± ·11 0·16 ± ·19	1·17 ± ·11 1·65 ± ·19	1·98 ± ·11 2·79 ± ·19
Kawamil (14·2) ...	19 Characters ... 6 Indices ...	0·34 ± ·11 0·16 ± ·19	— —	0·76 ± ·11 1·65 ± ·19	0·48 ± ·11 0·39 ± ·19
Beit-Allam (8·3) ...	19 Characters ... 6 Indices ...	1·17 ± ·11 1·65 ± ·19	0·76 ± ·11 1·65 ± ·19	— —	1·16 ± ·11 2·65 ± ·19
Naqada South (25·9)	19 Characters ... 6 Indices ...	1·98 ± ·11 2·79 ± ·19	0·48 ± ·11 0·39 ± ·19	1·16 ± ·11 2·65 ± ·19	— —

* The numbers in brackets following the designation of the races are the mean numbers of crania available for the characters used in computing the Coefficients of Racial Likeness.

Although the series are short, the lowness of the Coefficients in Table II may be taken to indicate very close affinity. The Kawamil series most closely represents the mean type and the El-Amrah skulls stand rather apart from the others. Ten characters—viz. 100 H'/L, L, H', LB, U, J, NB, O_1', O_2 and fmb—of the 19 compared show no significant differences between the four series. The values of α† for the other nine characters are:

		100 B/L	100 B/H'	100 NB/NH'	100 O_2/O_1'	100 fmb/fml	B	B'	NH'	fml
El-Amrah and	Kawamil	0·07	0·23	3·17	2·07	1·46	0·45	2·72	0·47	4·04
	Beit-Allam	3·66	2·47	1·86	1·32	5·59	2·71	6·86	1·77	10·29
	Naqada South	0·04	0·01	9·86	11·53	1·24	0·03	9·17	7·23	1·22
Kawamil and	Beit-Allam	6·62	5·09	0·08	0·30	2·36	7·34	1·70	0·68	2·76
	Naqada South	0·02	0·54	2·38	5·17	0·07	1·24	2·76	6·25	1·83
Beit-Allam and	Naqada South	7·17	3·47	2·48	4·44	3·38	2·60	0·01	1·57	8·20

The orbital is the only index which shows more than one value of α greater than five and the low value for the Naqada South series is responsible for the divergences in that case. The small frontal breadth (B') of the El-Amrah, the small nasal height (NH') of the Naqada South and the large foraminal length (fml) of the Beit-Allam series are also more or less distinctive. But by far the greater number of the cranial characters, and in particular those of the calvaria, are so similar for the four series that they might represent identical types. It is reasonable to suppose that three of the series—the Kawamil, Beit-Allam and Naqada South—are quite pure representatives of the same race: their pooled means are given in Column 6 of Table I (p. 14). The El-Amrah skulls diverge slightly from that type. The standard deviations of the three pooled distributions may be compared with those of the long series of 26th to 30th Dynasties skulls which is dealt with in a later section of this paper.

---

† With the usual notation $\alpha = \frac{n_s n_s'}{n_s + n_s'}\left\{\frac{M_s - M_s'}{\sigma_s}\right\}$ and its value is taken to indicate a significant difference when it is greater than 5.

	L	B	$100\,B/L$	B'	$100\,NB/NH'$
Fouquet's Pre-dynastic Series	5·79 ± ·36	5·06 ± ·31	3·11 ± ·19	6·41 ± ·39	4·69 ± ·35
26th to 30th Dynasty Series .	5·72 ± ·09	4·76 ± ·08	2·68 ± ·04	4·05 ± ·06	3·82 ± ·06*

* This S.D. is not of $100\,NB/NH'$ but of the very similar index $100\,NB/NH, R$.

The population made up by combining Fouquet's three series is certainly less homogeneous than that of the 26th to 30th Dynasties, but a comparison of the individual distributions suggests that the heterogeneity is inherent in each of them and is not due to the pooling. Provisionally we may assume that the pooled means are those of a pure racial type and we may speak of that as the Egyptian Aeneolithic type*. It will be shown that this primitive Aeneolithic race is most clearly differentiated from all the racial types that inhabited Egypt in Dynastic times and it must have been exterminated altogether, or, as the cultural evidence suggests, absorbed in a much larger population. The type is dolichocephalic and among African races it is distinguished by its large calvaria. Many modern races equal it in skull length, but apparently none but primitive types have as great a skull base (nasio-basial length) ($LB = 105·5$). The Eskimo LB length is 104·9, the Aino 105·4 and the Moriori 105·6†. It is reasonable to suppose that this primitive Egyptian Aeneolithic race was directly descended from the people whose palaeoliths are found on the high terraces on either side of the Nile Valley.

At least one other race was living in Upper Egypt at the time when the Aeneolithic people were still distinct in type. In the well-known work of Thomson and MacIver (13) individual measurements of 1500 skulls are provided. The material is divided up into series dating from Early Pre-dynastic to Roman times. Unfortunately the measurements are few in number and many of them are not defined precisely although there can be no doubt that they correspond exactly to measurements that are ordinarily taken in the Biometric Laboratory.

The Pre-dynastic skulls are divided into two groups, an early and a late; the relative ages having been determined by the method of Petrie's sequence dates. Of the ♂ skulls, 20 of the Early Pre-dynastic came from Abydos, 10 from El-Amrah and 25 from Hou: the pooled means are shown in Column 9 of Table I. Twenty-eight of the Late Pre-dynastic ♂ specimens came from El-Amrah and 99 from Hou; their pooled means are in Column 10[*].

---

* Although a few small metal objects were found in the early graves opened by J. de Morgan, he supposed that they belonged to a Neolithic stage of culture.

† These values are taken from *Biometrika*, Vol. XIV. p. 222.

Several large Early Pre-dynastic cemeteries were excavated by Petrie and Quibell in 1895 and an account of the work was given in the following year[†](3). Fawcett's paper of 1901 (6) dealt with 300 of these skulls. They had been taken from at least four distinct cemeteries within a radius of three or four miles, but the grave furniture found at the different sites was so similar that the fact that a homogeneous population was being dealt with appeared to be conclusively established. An examination of the skull measurements led to the same conclusion. The standard deviations of characters are quite as low as for populations which craniologists accept as homogeneous, though the majority of them are slightly greater than those of a long series—the *E* series—of Egyptian skulls of the 26th to 30th Dynasties which will be dealt with in a later section of this paper[‡]. A comparison of the mean measurements of the series from different cemeteries suggested, however, that the divergences, though small, might be of significance when considered in relation to the quantity of material that has become available since Fawcett wrote. By far the greater number of the so-called Naqada crania came from two cemeteries: one near the village of Nubt and the other two miles to the north towards Ballas[§]. These form the first series—the skull numbers not preceded by a letter and which for convenience we will term the *A* series—and the *Q* series respectively in Fawcett's tables of individual measurements. The *T* and *B* series were from two distinct burial groups within half a mile of the first site. The mean measurements of the *A* and *Q* series are almost identical and their combined means[||] are shown in Column 7 of Table I. The *T*, *B* and *R* series are very similar to one another and they agree with the *A* and *Q* series for the majority of characters but are yet sufficiently remote for others to make a further enquiry desirable. The splitting up of a relatively short series of crania into groups which are supposed more or less racially distinct, when the whole material came from a single district and was homogeneous from the standpoint of the archaeological evidence, is a practice that the biometric craniologist is bound to discredit unless very strong evidence can be adduced to support it. There are certainly not two *distinct* types among the Naqada crania. There were two distinct racial types in early Egypt, but they were quite as similar to one another as adjacent racial populations are generally found to be. Nearly all our cranial series represent intermediate types between those two and the differences between them are very small; so small that with less abundant material it would have been impossible to say that they had a real statistical significance. Our contention is that Fawcett's Naqada series of skulls contains representatives of two slightly differing populations which are both intermediate between the two extreme Egyptian types, and the evidence in favour of that view is partly deduced from a comparison of the two groups themselves, but its real justification is found in comparing their affinities with other early series. The Coefficients between the Naqada groups and Thomson and MacIver's Pre-dynastic series are given in the following Table:

---

[*] In the earlier work of MacIver (12) measurements of smaller numbers of Pre-dynastic crania were provided. The means of these—taking Early and Late periods together—were reduced by Miss M. A. Lewenz and Professor Karl Pearson and quoted by Fawcett ((6), p. 426). Individual measurements of 33 of the 44 skulls were given again in (13), so the earlier series is no longer of any importance; it is obviously intermediate between our Early and Late Pre-dynastic types. MacIver gave orbital measurements that were not reproduced in (13), but when divided up the numbers become so small that I have omitted to consider them.

[†] It was orignally thought that the Naqada people were of Early Dynastic date but further evidence showed that they must be assigned to the Pre-dynastic period.

[‡] The standard deviations of the Naqada, the *E* and three other series are compared in *Biometrika*, Vol. xiv. p. 208.

[§] The "New Race" cemeteries excavated by Petrie and Quibell were four or five miles north of the village of Naqada. The sites are shown in (3) Plates I and II. J. de Morgan worked in Naqada itself and it was there that he found Menes and the Naqada South skulls.

[||] With the modifications of Fawcett's measurements explained in the notes to the Table itself.

TABLE III.

Coefficients of Racial Likeness between some Pre-dynastic Series.

		Early Pre-dynastic (40·9)*	Naqada *A* and *Q* Series (66·1)	Late Pre-dynastic (105·6)	Naqada *B*, *T* and *R* Series (18·6)
Early Pre-dynastic (40·9)	14 Characters ... 6 Indices and Angles	— —	0·31 ± ·13 0·53 ± ·19	2·17 ± ·13 2·09 ± ·19	3·55 ± ·13 6·04 ± ·19
Naqada *A* and *Q* Series (66·1)	14 Characters ... 6 Indices and Angles	0·31 ± ·13 0·53 ± ·19	- –	0·73 ± ·13 1·62 ± ·19	1·62 ± ·13† 3·25 ± ·19
Late Pre-dynastic (105·6)	14 Characters ... 6 Indices and Angles	2·17 ± ·13 2·09 ± ·19	0·73 ± ·13 1·62 ± ·19	—	1·66 ± ·13 3·40 ± ·19
Naqada *B*, *T* and *R* Series (18·6)	14 Characters ... 6 Indices and Angles	3·55 ± ·13 6·04 ± ·19	1·62 ± ·13† 3·25 ± ·19	1·66 ± ·13 3·40 ± ·19	— —

* See footnote to Table II, p. 16.

† These Coefficients are based on 14 characters and 6 indices and angles so that they may be comparable with the others in the table. When they are calculated for the full 31 characters, of which 12 are indices and angles, they become 0·65 ± ·09 and 1·54 ± ·14, so they are substantially reduced. The added characters show no significant differences.

The low order of the Coefficients in Table III shows that we are dealing with closely related types. In comparing them it is important to pay due regard to the number of skulls available for each series—the number in brackets following the designation of the race. The Coefficients involving the Naqada *B*, *T* and *R* series are all symptomatic of greater racial divergences than those between any pair of the other three types, while the former stands closest to the Late Pre-dynastic. The Naqada *A* and *Q* group is almost identically the same as the Early Pre-dynastic, but between the latter and the Naqada *B*, *T* and *R* group there is a difference in type on which considerable stress may be laid as the numbers of skulls in the two series are small. The Coefficients arrange the series in the order shown in the table and that order is a perfectly consistent one. Turning to a comparison of the individual characters we find that 8 of the 14 compared—viz. $100\,NB/NH'$, L, H', LB, $G'H$, NH', NB and $A\angle$—show no significant differences whatever. The only values of α greater than 5 are for the characters shown below.

		B	$100\,B/L$	$100\,B/H'$	J	$100\,H'/L$	$N\angle$
Early Pre-dynastic and	Naqada *A* and *Q* ...	2·45	0·07	1·63	3·93	2·39	4·14
	Late Pre-dynastic...	6·55	0·82	9·06	12·74	7·53	0·11
	Naqada *B*, *T* and *R*	15·13	11·54	18·79	1·80	2·82	7·37
Naqada *A* and *Q* and	Late Pre-dynastic ...	1·14	0·62	3·85	1·51	1·76	4·81
	Naqada *B*, *T* and *R*...	8·96	11·95	11·18	0·04	0·42	1·56
Late Pre-dynastic and	Naqada *B*, *T* and *R*	6·00	9·47	5·21	0·40	0·02	7·83

The most significant differences are between the calvarial and zygomatic breadths and the indices $100\,B/L$ and $100\,B/H'$ which vary with B while H' and L remain constant. No stress can be laid on the single significant difference between the indices $100\,H'/L$. The two values of α greater than 5 in the case of the nasal angle are of greater significance. The actual measurements are:

	B	$100\,B/L$	$100\,B/H'$	J	H'	$N\angle$
Early Pre-dynastic	131·4	71·7	98·1	124·3	134·0	67°·0
Naqada *A* and *Q*...	132·7	71·8	99·2	125·9	133·8	65°·5
Late Pre-dynastic	133·4	72·1	100·5	127·0	132·7	66°·8
Naqada *B*, *T* and *R*	136·1	74·0	103·0	126·2	132·2	64°·3

The gradation shown in the case of the first three of these characters confirms the reasonableness of the order suggested by the Coefficients. H', though showing no significant differences, exhibits the same order. The discordance in the sequence of zygomatic breadths may well be attributed to the errors of random sampling. The variability of the nasal angles would appear to be due to some other cause and it is almost certainly the result of slight differences between the methods of measuring the sides of the fundamental triangle employed by the workers concerned*. The differences between the two Naqada groups are precisely similar to those which distinguish them from Thomson and MacIver's two series. From Early to Late Pre-dynastic times—a period of at least 2000 years—the change in the skull form of the population of Upper Egypt was exceedingly small. It may have been due to slow, progressive evolution or to slight admixture with another strain. A comparison of these Pre-dynastic types with Fouquet's Aeneolithic will be of interest. The Coefficients are:

---

* Probably due to differences in the determination of the alveolar point: see Note p. 58.

TABLE IV.

Coefficients of Racial Likeness between Fouquet's Aeneolithic and other Pre-dynastic Series.

		Early Pre-dynastic (40·9)	Naqada *A* and *Q* (66·1)	Late Pre-dynastic (105·6)	Naqada *B*, *T* and *R* (18·6)
Aeneolithic (48·7)	All Characters ...	11·23 ± ·14	9·18 ± ·11	10·71 ± ·14	5·28 ± ·11
	Number of Characters	11	19	11	19
	Indices and Angles ...	1·31 ± ·24	0·29 ± ·19	1·66 ± ·24	1·41 ± ·19
	Number of Characters	4	6	4	6

The most striking feature of Table IV is the great difference between the values of the Coefficients for all characters and for indices and angles only. With Thomson and MacIver's and Fouquet's series the only four indices that can be compared are $100\,B/L$, $100\,H'/L$, $100\,B/H'$ and $100\,NB/NH'$, while two others—$100\,O_2/O_1'$ and $100\,fmb/fml$—take part in the comparison with the Naqada groups. Only one difference of indices is significant; that is with the Naqada *B*, *T* and *R* series where $\alpha = 6{\cdot}87$ for $100\,B/H'$. The high values of the Coefficients for all characters are due entirely to the greater size of the Aeneolithic calvaria: facial direct measurements show no significant differences whatever. So the Aeneolithic type was almost identical with the Pre-dynastic Egyptian in size and shape of face and shape of calvaria, but it greatly exceeded the more advanced type in absolute size of calvaria*. The relation is a curious one and we can offer no explanation of it, but must consider the types to be perfectly distinct. The Aeneolithic people may have been absorbed into the larger population, but there is not the slightest suggestion that the latter was thereby modified in any way.

In the last column of Table I the mean measurements of a series of Pre-dynastic skulls discovered at El-Khozan and measured by E. Chantre are quoted from (15), p. 57. The measurements are presented in such an altogether unsatisfactory way that it is not possible to determine the relationships of the type with any certainty†. From a study of the grave finds Chantre supposed that the skulls belonged to the same race and were contemporary with J. de Morgan's series (see (14)) and he said that this conclusion was confirmed by the similarity of the skulls themselves. But the mean measurements are obviously very dissimilar to those of Fouquet's Aeneolithic type. They apparently resemble the Late Pre-dynastic type more than any other. No stress can be laid on the significance of the small *H'* and *J* of the El-Khozan skulls as either mean may easily have been based on no more than half of the 24 specimens.

---

* Judging from the drawings provided by Fouquet in appendices to J. de Morgan's volumes (10) and (11) the skulls of the Aeneolithic type had well-marked superciliary ridges, mastoids and inionic prominences and were decidedly more muscular than dynastic Egyptian crania.

† See notes to Table I.

A comparison of all the available skull measurements leads to the conclusion that in Pre-dynastic times there were two distinct racial types in Upper Egypt. The first and more primitive was characterised by a large calvaria and it was probably represented by a small and dwindling population. The second, which was carried down to historic times, was being transformed by some slow process of evolution or racial admixture. Both main types had low cephalic indices which distinguish them from the type of all later Egyptians. The apparent relationships of the various series are represented schematically in Fig. 1, p. 2.

4. *Skulls of the First to the Sixth Dynasties.* The first six dynasties cover the protodynastic era and the Old Kingdom. At the beginning of that period a change in the administration of the country took place which was likely to be of considerable ethnic importance. The kingdoms of Upper and Lower Egypt had been ruled by one king before the accession of Menes, the founder of the First Dynasty, but he unified the country and facilitated the intercourse between the north and the south.

All the available mean measurements of Early Dynastic series of skulls are shown in Table V. The series in the third and later columns is one which was measured by Mr G. H. Motley and the means are now given for the first time*. The collection consists of 48 skulls excavated by Sir Flinders Petrie in 1922 and presented by him to the Biometric Laboratory. They belonged to court-officials of the reigns of Zev and Zet, third and fourth kings of the First Dynasty. The bodies were buried in a square, the graves forming the border or sides, at Abydos between the Shuneh and the Coptic Deir. Each king had a fresh square of burials†. The full account of these excavations will be given in a volume entitled *Tombs of the Courtiers*, to be published shortly by Sir Flinders Petrie.

Of the 48 skulls one (No. 443) was almost certainly that of a negro; it is possible that some of the others showed, but far less definitely, negroid characters. Of the remaining 47 crania, 37 were allotted to the male, 10 to the female sex. Actually 81 crania were obtained and in sufficiently good condition to be treated with paraffin preparatory to transport; through some misadventure only 48 were packed and reached England. A series of 15 measurements were taken by Sir Flinders Petrie in the field; seven of these, our L (the maximum length), B (maximum breadth), H' (the basio-bregmatic height), LB (the basio-nasal height), $G'H$ (upper face height, nasion to alveolar point), GL (profile length from basion to alveolar point) and J (the bizygomatic breadth), are customary cranial measurements. The remaining eight measurements were taken for comparison with statuary, and it is less certain that they are: *Orbital Breadth* (? our O_1), *Jaw Breadth* (? our condylar width W_1), *Facial Height* (Nasion to base of jaw), *Jaw Height* (gnathion to infradental + height of teeth), *Nasidental Height* (nasion to line of teeth), *Nasal Height* (? our NH'), *Internal Orbital Width* and *External Orbital Width*, both measured

---

* Measurements of the individual skulls and remarks on cranial anomalies will be found in the Appendix to this paper.

† There is reason to believe that at least some of the interments must, by the attitude of the skeletons, have been cases of persons buried alive. An entire skeleton, to illustrate this, was brought back by Sir Flinders Petrie and is now mounted in the Museum of the Laboratory.

inside the orbits. As we invariably measure the external and internal ocular distances on the living subject, it seems highly important to include these among craniological measurements. I am inclined to think that the external orbital distance is better not taken *inside* the orbit, as then for comparison with the living we must allow for thickness of flesh; it would, I suggest, be advisable to take it from the inside border of one orbital ridge to that of the other, the measurement to be taken at the fronto-malar sutures, which can be closely approximated to in the living. The terminals of the internal orbital measurement are less easy to standardise. We have already two craniological measurements of this kind, namely *DC* the dacryal chord, and *SC* the minimum chord between the two naso-maxillary sutures, the so-called simotic chord. There is no doubt that the dacryal chord would better represent the internal orbital distance, and might be adopted for this purpose to save multiplication of cranial measurements*. If we add to the dacryal chord the widths of the two orbits we have for the mean total $\begin{Bmatrix} \text{♀}\,21·56 + 40·65 + 40·09 = 102·30 \\ \text{♂}\,22·82 + 42·77 + 42·33 = 107·92 \end{Bmatrix}$. But this is not a true measurement of the external ocular distance as the three segments are not in the same straight line. Using the measurement indicated above we find that the external ocular distance for males = 98·4 and for females 93·2 mms., while Sir Flinders Petrie's measurements for these 37 and 10 crania are 96·4 and 92·1 mm., so that taking the measurements "inside the orbits" makes a difference of 2 mm. in the male and 1 mm. in the female. The measurement from one fronto-malar to the other fronto-malar point seems to be somewhat in excess of the "inside the orbits" external ocular distance. It has the advantage, however, of being both an easily determined length on the skull and also a readily measured quantity on the living head. On the crania we have compared, Sir Flinders Petrie's field measurements, when taken for the same characters as those used by us, agree reasonably, 0·5 to 1 mm. being the usual difference†. A much greater divergence occurred in the sexing. We were at first much inclined to make the whole series *male*, but have finally admitted ten to be female. There is no archaeological evidence of sex, and failing this, we feel compelled to retain as a minimum four-fifths of the material as male. It is of course difficult with a short series of skulls, where we are not compelled by the circumstances of the interment—very peculiar in this case—to suppose both sexes represented, or represented in approximately equal numbers, to draw a marked line anatomically across the series separating it into two sex groups. It was indeed not hard to classify the material into two categories, and these might have been regarded as a sex-division; but they appeared to us to correspond far more to racial than to sexual differences. In other words we doubt the homogeneity of our material, we believe it to consist of a mixture of races, or better a race with individuals of more or less mixed blood.

---

* Our value of the dacryal chord for 34 ♂'s is 22·8 mm. Sir Flinders Petrie's internal orbital width, measured "inside the orbits," is 25·6 mm. for the same crania. This would only give a difference of 0·7 mm. in determining the distance of the pupillary centre from the sagittal plane and would serve for portraiture.

† The field measurements were taken before paraffining, but that process seems to have made no appreciable difference to the skulls.

TABLE V.

*Mean Measurements of Egyptian Series from the First to the Fifth Dynasties. Males**.

Dynasty ...	1st Private Tombs	1st Royal Tombs	1st and 2nd Royal Tombs	4th and 5th	4th
Locality ...	Abydos	Abydos	Abydos	Deshasheh and Medum	Sakkarah
Measured by...	Thomson and MacIver	Motley	Thomson and MacIver	Thomson and MacIver	Broca
F	182·4 (43)	182·3 (36)	182·2 (27)	183·0 (30)	—
L	183·8 (43)	184·9 (36)	184·8 (27)	184·9 (54)	185 (31)
B	133·7 (41)	136·2 (37)	137·7 (26)	139·3 (54)	141 (31)
B'	—	93·0 (35)	—	—	95 (31)
H'	135·1 (38)	132·1 (35)	132·8 (24)	136·0 (50)	137 (31)
LB	102·3 (30)	102·0 (33)	100·1 (19)	101·8 (47)	—
U	—	513·5 (30)	—	—	518 (31)
G'H	70·6 (29)	74·7 (31)	72·2 (20)	71·3 (19)	—
J	130·9 (25)	129·7 (31)	129·9 (16)	127·1 (16)	131 (31)
NH'	51·4 (30)	53·7 (32)	51·9 (21)	52·0 (44)	52 (31)
NB	25·8 (30)	25·5 (33)	25·2 (21)	25·1 (43)	24 (31)
DC	—	22·8 (34)	—	—	24 (31)
$O_1'R$	—	40·0 (32)	—	—	38 (31)†
O_2R	—	33·5 (33)	—	—	33 (31)†
GL	98·6 (29)	96·2 (30)	96·5 (19)	96·3 (42)	—
fml	—	37·0 (33)	—	—	36 (31)
fmb	—	30·0 (32)	—	—	30 (31)
100 *B/L*	{72·7 (41)}	73·8 (36)	{74·9 (26)}	{75·4 (50)}	{76·2 (31)}
100 *H'/L*	{73·6 (38)}	71·3 (34)	{72·0 (23)}	{72·6 (50)}	{74·1 (31)}
100 *B/H'*	{99·0 (38)}	103·6 (35)	{103·7 (24)}	{102·4 (50)}	{102·9 (31)}
100 (*B* − *H'*)/*L*	{−0·9 (38)}	2·64 (34)	{2·9 (23)}	{1·8 (50)}	{2·2 (31)}
100 *NB/NH'*	{50·4 (30)}	47·9 (32)	{48·5 (20)}	{48·2 (44)}	{46·1 (31)}
100 O_2/O_1', *R*	—	86·1 (32)	—	—	{86·8 (31)}†
100 *fmb/fml*	—	81·5 (32)	—	—	{83·3 (31)}
N ∠	{66°·2 (29)}	63°·4 (30)	{65°·5 (19)}	{64°·6 (19)}	—
A ∠	{72°·8 (29)}	72°·4 (30)	{71°·5 (19)}	{73°·4 (19)}	—
B ∠	{41°·0 (29)}	44°·2 (30)	{43°·0 (19)}	{42°·0 (19)}	—

Dynasty	1st Royal Tombs
Locality	Abydos
Measured by	Motley
Capacity	1364·3 (24)
W (grs.)	668·7 (32)
L'	185·5 (31)
H	133·8 (30)
OH	110·3 (31)
Q	304·9 (31)
Q'	305·5 (30)
S	369·0 (29)
S_1	123·7 (32)
S_2	129·4 (34)
S_3	116·9 (32)
PH	21·3 (34)
GB	97·3 (33)
NH, *R*	53·9 (32)
NH, *L*	54·0 (32)
DS	12·2 (29)
DA	35·8 (29)
SS	4·58 (30)
SC	10·97 (34)
O_1R	42·8 (33)
O_1L	42·3 (31)
O_2R	33·5 (33)
O_2L	33·9 (31)
EOW	98·4 (37)
G_1	54·6 (32)

Dynasty	1st Royal Tombs
Locality	Abydos
Measured by	Motley
G_1'	49·5 (32)
G_2	42·7 (29)
EB	42·8 (29)
EH	14·7 (29)
100 *B/L'*	73·4 (31)
100 *H/L'*	72·1 (30)
100 *H/L*	72·3 (30)
100 *B/H*	102·1 (30)
100 *G'H/GB*	76·8 (30)
100 *NB/NH*, *R*	47·5 (31)
100 *NB/NH*, *L*	47·7 (32)
100 O_2/O_1, *R*	78·2 (33)
100 O_2/O_1, *L*	80·4 (31)
100 *DS/DC*	32·6 (29)
100 *SS/SC*	42·4 (30)
100 G_2/G_1	78·6 (29)
100 G_2/G_1'	{86·4 (29)}
100 *EH/EB*	34·4 (29)
P ∠	86°·1 (32)
θ_1	30°·3 (30)
θ_2	13°·9 (30)
S_3'	96·0 (32)
Oc. I.	59·9 (32)

---

* The angles and indices in curled brackets were found from the mean measurements of the component lengths instead of from the individual measurements.

† These measurements may be of either the right or left orbit.

Without venturing to be in the least dogmatic, on such slender evidence, we should expect a divergence from the prehistoric Egyptians by their mixture with a race of a coarser, stronger type. If the line drawn across the material corresponds to more and less of this stronger coarser race, it will be obvious that there is danger in such a small series of confusing such a racial line with a sex line. Were anyone to assert on other grounds than we have had before us that the whole series is male, there are certainly only a very few crania about which we should be inclined to contest his assertion.

Measurements of two other series of First Dynasty skulls have been furnished by Thomson and MacIver (13). One came from a private graveyard and the other from the Royal Tombs of the First and Second Dynasties; both were from Abydos. The same writers have provided measurements of Fourth and Fifth Dynasties skulls from Deshasheh and Medum in Lower Egypt. The mean measurements are shown in Column 5 of Table V.

The Coefficients of Racial Likeness between the above series were calculated. The numbers of skulls are fairly adequate but unfortunately there are no more than 14 characters, of which 6 are indices and angles, that can be used for that purpose.

TABLE VI.

Coefficients of Racial Likeness between Early Dynastic Series.

		1st Dynasty Private Tombs (33·6)	1st Dynasty Royal Tombs (Motley) (33·2)	1st and 2nd Dynasties Royal Tombs (21·8)	4th and 5th Dynasties (39·9)
1st Dynasty ... Private Tombs (33·6)	14 Characters ... 6 Indices and Angles	— —	5·37 ± ·13 7·72 ± ·19	3·54 ± ·13 5·28 ± ·19	5·59 ± ·13 6·97 ± ·19
1st Dynasty ... Royal Tombs (33·2) ...	14 Characters ... 6 Indices and Angles	5·37 ± ·13 7·72 ± ·19	— —	0·66 ± ·13 0·36 ± ·19	2·80 ± ·13 1·62 ± ·19
1st and 2nd Dynasties Royal Tombs (21·8) ...	14 Characters ... 6 Indices and Angles	3·54 ± ·13 5·28 ± ·19	0·66 ± ·13 0·36 ± ·19	— —	0·39 ± ·13 0·07 ± ·19
4th and 5th Dynasties (39·9)	14 Characters ... 6 Indices and Angles	5·59 ± ·13 6·97 ± ·19	2·80 ± ·13 1·62 ± ·19	0·39 ± ·13 0·07 ± ·19	— —

The most striking feature of this table is the highness of the Coefficients between the First Dynasty skulls from the private tombs and the other series. The numbers of skulls being small, the Coefficients are of an order which must be taken to indicate perfectly distinct racial types. The two sets of skulls from the royal cemeteries are very similar to one another and are far more closely related to the later type from Lower Egypt than to the contemporary civil population of their own town. The Coefficient between the First and Second Dynasties series and the Fourth and Fifth Dynasties series is as low as it might have been if the two had been drawn from the same population. Motley's series, however, is distinct from that of the Fourth and Fifth Dynasties. From this evidence alone, which by itself would be very scanty of course, we may suppose that the courtier population of Abydos in the First and Second Dynasties was intrusive and that it came from Lower Egypt where the type persisted until at least as late as the Fifth Dynasty. A comparison of the individual characters of the four series can be made in the usual way by considering the values of the α's between them. There are only four characters—viz. L, LB, NB and $A\angle$—which show no values of α greater than 5. The α's for the remaining ten are given in the table below.

All the greatest values of α are in the first three rows of this table. The indices $100\,B/H'$ and $100\,B/L$ of the First Dynasty Private Tombs series are quite distinct from the values for the other three types and the divergences are evidently due primarily to differences in calvarial breadth. In the last rows there are only five characters with values of α greater than 5. In the whole table the only significant differences of $G'H$ and NH' are those from Motley's First Dynasty series. The mean facial and nasal heights of those skulls are greater, and in the majority of cases, quite significantly greater than those of any other ancient Egyptian series; and other characters, not provided by Thomson and MacIver, also distinguish them, so there is clear evidence of admixture with some alien strain. At the same time they stand very close to the other Royal Tombs series of the First and Second Dynasties and they are primarily of Egyptian type. The Fourth and Fifth Dynasties series from Lower Egypt is distinguished from the royal retainers of the First and Second Dynasties by having a much greater basio-bregmatic height.

		$100\ B/H'$	$100\ B/L$	B	$100\ H'/L$	$100\ NB/NH'$	H'	$G'H$	NH'	J	$N\angle$
1st Dynasty Private Tombs	1st Dynasty Royal Tombs (Motley)	20·84	3·26	5·35	10·98	6·51	5·01	14·64	9·60	0·95	10·54
	1st and 2nd Dynasties Royal Tombs	17·56	10·80	11·22	4·24	2·92	2·38	1·76	0·36	4·68	0·52
	4th and 5th Dynasties	13·49	23·02	32·18	2·50	5·80	0·51	0·32	0·75	6·74	2·69
1st Dynasty Royal Tombs (Motley)	1st and 2nd Dynasties Royal Tombs	0·01	2·60	1·51	0·78	0·31	0·21	4·41	4·82	0·02	4·68
	4th and 5th Dynasties	1·61	7·51	9·29	3·97	0·11	9·57	7·92	6·28	3·41	1·52
1st and 2nd Dynasties Royal Tombs	4th and 5th Dynasties	1·48	0·60	1·98	0·66	0·08	5·08	0·46	0·01	3·00	0·70

The Coefficients between these Early Dynastic and the four Pre-dynastic series previously considered are given in Table VII.

TABLE VII.

Coefficients of Racial Likeness between Early Dynastic and Pre-dynastic Series.

		1st Dynasty Private Tombs (33·6)	1st Dynasty Royal Tombs Motley (33·2)	1st and 2nd Dynasties Royal Tombs (21·8)	4th and 5th Dynasties (39·9)
Early Pre-dynastic (40·9)	14 Characters ...	2·50 ± ·13	12·15 ± ·13	7·64 ± ·13	11·86 ± ·13
	6 Indices and Angles	−0·11 ± ·19	12·17 ± ·19	8·62 ± ·19	13·36 ± ·19
Naqada *A* and *Q* (66·1)	14 Characters ...	1·92 ± ·13	9·68 ± ·13*	5·98 ± ·13	10·79 ± ·13
	6 Indices and Angles	0·46 ± ·19	8·60 ± ·19	7·81 ± ·19	12·45 ± ·19
Late Pre-dynastic (105·6)	14 Characters ...	2·17 ± ·13	9·82 ± ·13	4·88 ± ·13	11·42 ± ·13
	6 Indices and Angles	1·84 ± ·19	9·99 ± ·19	6·55 ± ·19	13·75 ± ·19
Naqada *B*, *T* and *R* (18·6)	14 Characters ...	2·45 ± ·13	2·59 ± ·13*	0·56 ± ·13	1·27 ± ·13
	6 Indices and Angles	2·78 ± ·19	0·22 ± ·19	0·37 ± ·19	0·33 ± ·19

* The Coefficients between Motley's First Dynasty series and the two Naqada can also be calculated for the whole 31 characters, of which 12 are indices and angles; with the *A* and *Q* they are 6·36 ± ·09 and 6·65 ± ·14 and with the *B*, *T* and *R* 2·37 ± ·09 and 0·90 ± ·14 respectively for all characters and for indices and angles only.

The Coefficients in Table VII show that there is a distinct line of cleavage dividing the series into two opposed groups. The Private Tombs series of the First Dynasty belongs clearly to the Pre-dynastic stock and it is apparently most closely related to the Late Pre-dynastic type which is represented by more skulls than the other two. And those four series are most clearly differentiated from the other three arranged horizontally in the table; the latter form another group of closely inter-related types. For convenience we may term these groups the Upper—the Pre-dynastic—and Lower Egyptian types respectively; the appropriateness of the titles will be more apparent when later series are considered. The Naqada *A* and *Q* group behaves in every way like Thomson and MacIver's Pre-dynastic types but the distinctiveness of the *B*, *T* and *R* skulls is again brought out in Table VII. The series being a short one its Coefficients are for the most part of a lower order than the others in the table but it clearly inclines more to the Lower than to the Upper Egyptian type and may be considered intermediate between the two. The division between the two main groups is quite sharp, but there is very definite evidence to show that the transformation from one to the other had started in prehistoric times. A comparison of all the Coefficients shows that the Early Pre-dynastic is the series most widely removed from the Lower Egyptian type and the Naqada *A* and *Q* group is more distant than the Late Pre-dynastic. A similar gradation, providing the principal justification for the arrangement shown in Fig. 1, p. 2, is found among the Lower Egyptian series. The relationships are brought out more clearly by comparing individual characters. Whether comparison is made between series within either group or between series of different groups, the same characters always show the most significant differences and the same are constant throughout so long as Motley's First Dynasty series is not concerned in the comparison. From the latter, greatest divergences—i.e. the highest values of α—are in every case for the characters $G'H$ and NH', but otherwise it behaves like a pure Egyptian series*. In general the relationships between these series are such that:

(i) L, LB, NB and $A\angle$ show no significant differences whatever,

(ii) $100\,H'/L$, $G'H$ and NH' are practically constant,

(iii) $N\angle$, H' and $100\,NB/NH'$ have small but orderly differences, and

(iv) the greatest differences are found always for B, $100\,B/H'$, $100\,B/L$ and to a less marked extent for J.

The measurements showing most significant differences are compared below:

		B	$100\,B/L$	$100\,B/H'$	J	H'	$100\,NB/NH'$	$N\angle$
Upper Egyptian Type	Early Pre-dynastic	131·4	71·7	98·1	124·3	134·0	50·3	67°·0
	Naqada *A* and *Q*	132·7	71·8	99·2	125·9	133·8	50·0	65°·5
	Late Pre-dynastic	133·4	72·1	100·5	127·0	132·7	51·0	66°·8
	1st Dynasty Private Tombs	133·7	72·7	99·0	130·9†	135·1	50·4	66°·2
	Naqada *B*, *T* and *R*	136·1	74·0	103·0	126·2	132·2	49·9	64°·3
Lower Egyptian Type	1st Dynasty Royal Tombs (Motley)	136·2	73·8	103·6	129·7	132·1	47·9	63°·4
	1st and 2nd Dynasties Royal Tombs	137·7	74·9	103·7	129·9	132·8	48·5	65°·5
	4th Dynasty (Broca)	137	76·2	102·9	131	137	46·1	—
	4th and 5th Dynasties	139·3	75·4	102·5	127·1†	136·0	48·2	64°·6

† These bizygomatic breadths are based on fewer crania than the corresponding calvarial breadths. In the case of the First Dynasty series B is the mean of 41 skulls and J of 25. The mean B for those 25 is 134·8 so the J for the 41 would almost certainly have been greater than the recorded value and more in accordance with those of the allied series. In the same way it is probable that the Fourth and Fifth Dynasties J should be greater than 127·1.

The most important differences between the Upper and Lower Egyptian types are those associated with the breadth of the calvaria—B, $100\,B/L$, and $100\,B/H'$; the latter is the broader and less dolichocephalic skull. Those differences are correlated with lesser differences between the zygomatic breadths. The Lower Egyptian type also has the lesser nasal index and nasal angle. The basio-bregmatic heights are less regular. The values of H' for the Fourth Dynasty series are quite significantly greater than those for the First Dynasty Royal Tombs skulls and it will be shown later that the smaller values are those of the pure type while the others can only be accounted for on the supposition that the skulls represent a slight admixture with a race foreign to Egypt. So we may conclude that the Upper Egyptian type has a greater basio-bregmatic height than the Lower. The Fourth Dynasty series from Sakkarah, of which measurements are given above and in Column 6 of Table V, was measured by Broca. The measurements are in his unpublished note-books but the means have been quoted by several French craniologists*. Unfortunately they are given in whole mms. only and perhaps not to the nearest mm., while it is not known how many of the 31 skulls each mean is based on. The means are very similar to those of the contemporary series from Deshasheh and Medum measured by Thomson and MacIver†.

---

* We are only considering here the 14 characters measured by Thomson and MacIver which can be used in calculating the Coefficients of Racial Likeness. There are other characters which distinguish Motley's First Dynasty series from all purer Egyptian types (see below pp. 42 and 43).

It will be convenient to consider first two later series from Lower Egypt. There are no others from that region: the majority of the ancient Egyptian skulls that have been measured came from the Thebaid.

5. *Series from Lower Egypt later than the Fifth Dynasty.* A long series of crania belonging to the period from the 26th to 30th Dynasties and containing upwards of 1700 individuals has been measured in the Biometric Laboratory by Miss A. Davin and others, and a paper on this series has recently been published‡. The object of that paper was to determine authoritative values for the correlation of cranial characters. This object can only be obtained by the measurement of a long series. The present use of that series is only a by-product of that enquiry, which has itself no direct bearing on the present paper. The standard deviations found for the E series have, however, been used in this and in other recent craniological work as constants for Coefficients of Racial Likeness. This is done because, being deduced from large numbers of crania, the values are more likely to express the general variation of the human skull within a fairly homogeneous group than standard deviations deduced from the small series which are all it is feasible to obtain as samples in the case of many racial groups. The mean measurements of the 26th to 30th Dynasties skulls are given in Table XIV. The specimens came

---

* The means given in Table V were taken from (16), p. 507 and (15), p. 57.

† In the winter of 1924—5 the British School excavated sites at Qau, 30 miles south of Assiut, of the Fourth to the Ninth Dynasties. The human remains discovered are of particular interest because there are no records of others from Middle Egypt. Through the courtesy of Sir Flinders Petrie I am able to give here the means of his field measurements of some of the ♂ skulls. The series has been presented to the Biometric Laboratory and a full report on it will appear later. The means are: $L = 184{\cdot}7$ (31), $B = 139{\cdot}6$ (31), $H' = 136{\cdot}9$ (30), $J = 127{\cdot}7$ (28), $DC = 24{\cdot}6$ (30), $GL = 95{\cdot}5$ (29), $LB = 101{\cdot}2$ (30), $G'H = 70{\cdot}8$ (29), $NH'? = 51{\cdot}2$ (29), $NB = 25{\cdot}4$ (29), $100\,B/L = \{75{\cdot}6\ (31)\}$, $100\,H'/L = \{74{\cdot}1\ (30)\}$, $100\,B/H' = \{102{\cdot}0\ (30)\}$, $100\,NB/NH' = 49{\cdot}6$ (29), $N\angle = \{64^\circ{\cdot}5\ (29)\}$, $A\angle = \{78^\circ{\cdot}8\ (29)\}$, $B\angle = \{41^\circ{\cdot}7\ (29)\}$. If these measurements are compared with the Fourth and Fifth Dynasties series from Deshasheh and Medum provided by Thomson and MacIver (Table V above) the two will be seen to be remarkably similar. Fourteen characters can be used in calculating the Coefficient of Racial Likeness between them, giving C.L. = $-0{\cdot}50 \pm {\cdot}13$ and C.L. = $-0{\cdot}46 \pm {\cdot}19$ for 6 indices and angles. So, as far as we can tell, the populations were absolutely identical and they only differ from the pure Lower Egyptian type by having a greater basio-bregmatic height.

‡ *Biometrika*, Vol. XVI. pp. 237—239. The series is known as the E series and the ♂ standard deviations had previously been published in *Biometrika*, Vol. XIV. p. 208 and the ♀ in Vol. XVI. p. 14.

from a single cemetery south of Gizeh which was excavated by Sir Flinders Petrie (see (5), p. 29). In Table VIII the measurements of a short series from Sakkarah dating from Ptolemaic times are compared with two other series from Lower Egypt—the 26th to 30th Dynasties from Gizeh and Broca's Fourth Dynasty series from Sakkarah. The individual measurements of the Ptolemaic skulls are in Broca's unpublished note-books and the means are given by Chantre ((15), p. 131) and other French craniologists. They are presented in as unsatisfactory a way as the rest of that material (see footnote || to Table I, p. 15 above).

TABLE VIII.

*Mean Measurements of Lower Egyptian Series. Males**.

	L	B	H'	B'	J	DC
E Series, 26th to 30th Dynasties	185·3 (895)	138·9 (896)	132·4 (884)	94·8 (935)	128·7 (785)	—
4th Dynasty, Sakkarah (Broca) ...	185 (31)	141 (31)	137 (31)	95 (31)	131 (31)	24 (31)
Ptolemaic Period, Sakkarah (Broca)	181 (13)	139 (13)	136 (13)	97 (13)	126 (13)	23 (13)
	O_1'	O_2	NH'	NB	fml	fmb
E Series, 26th to 30th Dynasties	38·4 (886)	33·7 (888)	51·5 (898)	24·4 (893)	35·6 (894)	30·0 (905)
4th Dynasty, Sakkarah (Broca) ...	38 (31)	33 (31)	52 (31)	24 (31)	36 (31)	30 (31)
Ptolemaic Period, Sakkarah (Broca)	39 (13)	33 (13)	50 (13)	24 (13)	35 (13)	29 (13)
	$100\ B/L$	$100\ H'/L$	$100\ B/H'$	$100\ O_2/O_1'$	$100\ NB/NH'$	$100\ fmb/fml$
E Series, 26th to 30th Dynasties	75·1 (884)	71·5 (884)	104·9 (884)	87·8 (886)	47·4 (893)	84·5 (889)
4th Dynasty, Sakkarah (Broca) ...	76·2 (31)	74·1 (31)	102·9 (31)	86·8 (31)	46·1 (31)	83·3 (31)
Ptolemaic Period, Sakkarah (Broca)	76·8 (13)	75·1 (13)	102·2 (13)	84·6 (13)	48·0 (13)	82·9 (13)

* A full account of the measurements of the 26th to 30th Dynasties series is given in the footnote to Table XIV. The indices of Broca's two series in Table VIII were all calculated from the mean values of the component lengths.

The following Coefficients of Racial Likeness were found:

TABLE IX.

Coefficients of Racial Likeness with the E Series (26th to 30th Dynasties) from Gizeh.

	26th to 30th Dynasties (885·4)			
	All Characters		Indices and Angles	
	C. L.	Number of Characters	C. L.	Number of Characters
1st Dynasty Private Tombs (33·6) ...	16·04 ± ·13	14	24·79 ± ·19	6
1st and 2nd Dynasties Royal Tombs (21·8)	1·20 ± ·13	14	1·85 ± ·19	6
1st Dynasty Royal Tombs (Motley) (33·2)	9·75 ± ·09	31	4·25 ± ·14	12
4th Dynasty (Broca) (31·0)...	3·89 ± ·13	13	3·62 ± ·24	4
4th and 5th Dynasties (39·9)	3·43 ± ·13	14	3·44 ± ·19	6

The 26th to 30th Dynasties series contains such a large number of skulls that Coefficients between it and closely allied types may be as large as those which would indicate distinct racial differences when shorter series were being compared. All the Coefficients with the Upper Egyptian types (not given in Table IX) are of the same order as those with the First Dynasty Private Tombs series above. The 26th to 30th Dynasties skulls are of the Lower Egyptian type and they show strong affinity with all the representatives of that race except Motley's aberrant series. From the First to the 30th Dynasties the type had hardly changed at all and the latest series shows no progression on that of the Fourth and Fifth Dynasties. A comparison of individual characters confirms that conclusion. The values of the distinctive characters $100\,B/H'$, $100\,NB/NB'$ and $N\angle$ as well as its cephalic index and calvarial breadth all relegate the 26th to 30th Dynasties series to the Lower Egyptian group and there is no suggestion of there having been any admixture whatever between the more stationary population of the Fayum and the Upper Egyptian race. If the E series is intercalated between the First and Fourth Dynasties series we get a more or less orderly sequence for the characters B and $100\,B/L$ but the increase in calvarial breadth is apparently not associated with an increase in zygomatic breadth.

	B	$100\,B/L$	J	H'	$100\,B/H'$	$100\,H'/L$
1st Dynasty Royal Tombs (Motley) ...	136·2	73·8	129·7	132·1	103·6	71·3
1st and 2nd Dynasties Royal Tombs ...	137·7	74·8	129·9	132·8	103·7	72·0
26th to 30th Dynasties	138·9	75·1	128·7	132·4	104·9	71·5
4th Dynasty (Broca)	137	76·2	131	137	102·9	74·1
4th and 5th Dynasties	139·3	75·4	127·1	136·0	102·5	72·6
Ptolemaic Period (Broca)	139	76·8	126	136	102·2	75·1

The basio-bregmatic height (H') and the indices $100\,B/H'$ and $100\,H'/L$ affected by it are responsible for all the greatest values of α between the six series. There is a clear distinction for those characters between the upper and lower triplets. This is somewhat surprising but it cannot disturb our belief in the very close affinities of all the types. There is clear evidence to show that the pure Upper Egyptian type is represented by the 26th to 30th Dynasties series*, and hence we must suppose that the other has been modified by some process of evolution or racial admixture. As far as we can tell that process only affected the calvarial height so that the only significant differences between the Fourth and Fifth Dynasties series and the 26th to 30th are for H', $100\,B/H'$, $100\,H'/L$ and NB, the last difference being small. We may note that the skulls of the First Dynasty series, which represent the Lower Egyptian type in Upper Egypt, are not absolutely pure, but they incline slightly towards the Upper Egyptian type, as shown in Fig. 1. That arrangement is justified by a comparison of all the Coefficients.

* The transformed Upper Egyptian types incline towards the 26th to 30th Dynasties series for all characters.

6. *Upper Egyptian Series later than the Fifth Dynasty measured by Thomson and MacIver.* There are six of these series and the mean measurements, which have not previously been published, are provided in Table XI. The number of skulls in each is sufficiently large to give a reliable determination of type. From a cursory inspection of the figures it is clear that three of the series resemble one another closely and are very similar to the First Dynasty Private Tombs skulls. The Coefficients of Racial Likeness are given in Table X.

TABLE X.

Coefficients of Racial Likeness between Series of Early and Middle Dynastic Crania.

		1st Dynasty Private Tombs (33·6)	6th to 12th Dynasties (168·6)	12th to 15th Dynasties (65·9)	18th Dynasty Shekh Ali (40·0)
1st Dynasty ... ... Private Tombs (33·6)...	14 Characters ...	—	2·88 ± ·13	1·88 ± ·13	3·42 ± ·13
	6 Indices and Angles	—	−0·13 ± ·19	0·87 ± ·19	1·58 ± ·19
6th to 12th Dynasties (168·6)	14 Characters ...	2·88 ± ·13	—	2·54 ± ·13	2·87 ± ·13
	6 Indices and Angles	−0·13 ± ·19	—	2·84 ± ·19	3·21 ± ·19
12th to 15th Dynasties (65·9)	14 Characters ...	1·88 ± ·13	2·54 ± ·13	—	0·14 ± ·13
	6 Indices and Angles	0·87 ± ·19	2·84 ± ·19	—	−0·25 ± ·19
18th Dynasty Shekh Ali (40·0)	14 Characters ...	3·42 ± ·13	2·87 ± ·13	0·14 ± ·13	—
	6 Indices and Angles	1·58 ± ·19	3·21 ± ·19	−0·25 ± ·19	—

Paying due regard to the differing numbers of skulls in the series, we may conclude from the Coefficients that they are arranged by their relationships in a sequence which is identically the same as the chronological order. This suggests that the skull form was being gradually modified in a particular direction though the change from the First to the 18th Dynasties was quite slight. Significant differences are only shown for the characters $100\,NB/NH'$, J, H', NB, $G'H$ and $100\,H'/L$. The last three of these are characters which are constant for pure Egyptian types, but no real significance can be attached to their small and erratic differences in the case of the four Early Dynastic series. There is no gradation in calvarial breadths or cephalic indices, but $100\,B/H'$, $100\,N/H'$, H' and $N\angle$ all give orders which are almost precisely the same as the one determined from the Coefficients and the series are obviously links in the transition from the Upper to the pure Lower Egyptian type of the 26th to 30th Dynasties series.

	$100\,B/H'$	H'	$100\,NB/NH'$	$N\angle$*
1st Dynasty Private Tombs ...	99·0	135·1	50·4	66°·2
6th to 12th Dynasties ...	99·6	134·8	50·4	65°·7
12th to 15th Dynasties ...	100·6	133·0	48·8	65°·7
18th Dynasty (Shekh Ali) ...	101·1	132·4	48·0	65°·6
26th to 30th Dynasties ...	104·9	132·4	47·4	63°·9

---

* Part of the difference between the last series and the others for this character may be due to differences in the determination of the alveolar point. See p. 53 below.

TABLE XI.

Upper Egyptian Series later than the 5th Dynasty measured by Thomson and MacIver. Males.*

Dynasty	6th—12th	12th—15th	18th	18th	Ptolemaic Period	Roman Period
Locality	Dend·rah	Hou and Abydos	Shekh Ali	Abydos	Denderah	Denderah
F	181·4 (175)	181·3 (67)	180·7 (41)	183·2 (50)	181·9 (78)	179·3 (50)
L	182·5 (175)	183·0 (67)	182·7 (41)	185·8 (50)	183·5 (77)	181·4 (50)
B	134·3 (173)	133·8 (66)	133·8 (41)	137·3 (50)	135·3 (78)	136·4 (50)
H'	134·8 (171)	133·0 (66)	132·4 (41)	134·2 (50)	131·5 (77)	130·7 (50)
LB	100·8 (167)	100·6 (66)	100·4 (40)	101·5 (50)	100·7 (76)	99·9 (50)
$G'H$	69·9 (169)	71·0 (66)	69·0 (40)	71·8 (50)	71·0 (76)	70·5 (50)
J	124·5 (157)	126·4 (61)	125·1 (37)	128·9 (49)	128·6 (74)	127·5 (46)
NH'	51·2 (168)	51·2 (67)	50·8 (39)	52·9 (50)	52·1 (77)	51·6 (48)
NB	25·8 (167)	25·0 (67)	24·4 (39)	25·8 (50)	25·7 (77)	25·5 (48)
GL	96·5 (166)	96·7 (66)	95·8 (40)	95·5 (50)	95·0 (76)	94·2 (50)
$100\,B/L$	73·6 (173)	73·1 (66)	73·2 (41)	73·9 (50)	73·7 (77)	75·2 (50)
$100\,H'/L$	73·9 (171)	72·7 (66)	72·5 (41)	72·2 (50)	71·7 (77)	72·1 (50)
$100\,B/H'$	99·6 (171)	100·6 (66)	101·1 (41)	102·3 (50)	102·9 (77)	104·4 (50)
$100\,(B-H')/L$	−0·3 (171)	0·4 (66)	0·8 (41)	1·7 (50)	2·1 (77)	3·1 (50)
$100\,NB/NH'$	50·4 (167)	48·8 (67)	48·0 (39)	48·8 (50)	49·3 (77)	49·1 (48)
$N\angle$	65°·7 (166)	65°·7 (66)	65°·6 (40)	63°·9 (50)	64°·3 (76)	64°·3 (50)
$A\angle$	72°·8 (166)	72°·0 (66)	73°·2 (40)	73°·3 (50)	73°·3 (76)	73°·2 (50)
$B\angle$	41°·5 (166)	42°·3 (66)	41°·2 (40)	42°·8 (50)	42°·4 (76)	42°·5 (50)

* All the indices and angles in this table were calculated from the mean values of the component lengths.

TABLE XII.

Coefficients of Racial Likeness between series of the Lower Egyptian Type.

		18th Dynasty Abydos (49·9)	Ptolemaic* Period (76·6)	Roman Period (49·3)	1st and 2nd Dynasties Royal Tombs (21·8)	4th and 5th Dynasties (89·9)	26th to 30th Dynasties (885·4)
18th Dynasty Abydos (49·9)	14 Characters ...	—	0·89 ± ·13	2·92 ± ·13	0·16 ± ·13	0·87 ± ·13	5·88 ± ·13
	6 Indices and Angles	—	− 0·56 ± ·19	1·15 ± ·19	0·51 ± ·19	0·61 ± ·19	5·40 ± ·19
Ptolemaic Period (76·6)	14 Characters ...	0·89 ± ·13	—	0·72 ± ·13	0·64 ± ·13	4·26 ± ·13	10·05 ± ·13
	6 Indices and Angles	−0·56 ± ·19	—	1·31 ± ·19	0·93 ± ·19	2·00 ± ·19	8·68 ± ·19
Roman Period (49·3)	14 Characters ...	2·92 ± ·13	0·72 ± ·13	—	0·65 ± ·13	3·55 ± ·13	5·35 ± ·13
	6 Indices and Angles	1·15 ± ·19	1·31 ± ·19	—	0·10 ± ·19	0·48 ± ·19	2·19 ± ·19

* Between the short Naqada *B*, *T* and *R* Series (see above, p. 18) and Thomson and MacIver's Ptolemaic we find coefficients of 0·04 ± ·18 for 14 characters and − 0·88 ± ·19 for 6 indices and angles. It is possible that the skulls that make up the former series represented a population of almost pure Lower Egyptian type which was intrusive in Upper Egypt in early times and had not mixed with the native population there. That the two Naqada groups were almost certainly contemporaneous is shown by the similarity of their grave furniture.

A comparison of the Coefficients of Racial Likeness with the 26th to 30th Dynasties series confirms the conclusion already arrived at that as the dynasties advanced the population of Upper Egypt approximated more and more to the stable type of the Lower Egyptian people.

The series of which the means are shown in the last three columns of Table XI appear to be of almost pure Lower Egyptian type although they came from Upper Egypt. The Coefficients between them and with three other series to which they are closely related are given in Table XII. The Ptolemaic and 18th Dynasty Abydos series are almost identical, while the Roman is very similar to the former but distinct from the latter. All three are closer to the First and Second Dynasties Royal Tombs series, than to any other of the Lower Egyptian type and the skulls of the 18th Dynasty from Abydos are the only ones that show very close affinity with those of the Fourth and Fifth Dynasties from Deshasheh and Medum in Lower Egypt.

The Fourth and Fifth Dynasties series was shown to be of absolutely pure Lower Egyptian type for all characters except the basio-bregmatic height and the index $100\,B/H'$. The 18th Dynasty Abydos series, though still inclining slightly towards the Upper Egyptian stock, is also primarily of the Lower Egyptian type and it too is characterised by a large H'. We can only suppose that there was a very slight alien element in both those populations. We may intercalate them between the other series to give the more or less orderly sequences shown below for all characters except H' and B/H', but the arrangement in Fig. 1, p. 2, more accurately represents their relationships.

	B	100 B/L	100 B/H'	100 NB/NH'	H'
Ptolemaic Period	135·3	73·7	102·9	49·3	131·5
18th Dynasty Abydos	137·3	73·9	102·3	48·8	134·2
Roman Period	136·4	75·2	104·4	49·4	130·7
1st and 2nd Dynasties Royal Tombs ...	137·7	74·9	103·7	48·5	132·8
26th to 30th Dynasties	138·9	75·1	104·9	47·4	132·4
4th and 5th Dynasties	139·3	75·4	102·5	48·2	136·0

There is a considerable difference between the 18th Dynasty skulls from Abydos and the contemporary series from Shekh Ali, the Coefficients between the two being 4·87 ± ·13 for 14 characters and 0·73 ± ·19 for six indices and angles. The Fourth and Fifth Dynasties and the 18th Dynasty Abydos series are the only two measured by Thomson and MacIver which do not represent intermediate types between the two races which populated Upper and Lower Egypt in the earliest times of which we have record, without the slightest infusion, as far as we can tell, of racial elements foreign to Egypt.

7. *Upper Egyptian Series later than the Fifth Dynasty measured by Chantre and Broca**. This material has all the defects of Chantre's prehistoric series (see

---

* We are unable to make any use of two papers by Pruner-Bey (17) and Biasutti (18) that are frequently referred to. The total number of skulls dealt with in each is considerable, but when divided into geographical and time groups they become quite inadequate.

TABLE XIII.

Mean Measurements of Upper Egyptian Series later than the 5th Dynasty measured by Chantre and Broca.*

Period	11th Dynasty	18th Dynasty	18th Dynasty	26th Dynasty	Ptolemaic
Locality	Thebes (Drah abou'l Neggah)	Gournah	Thebes	Gournah	Gournah
Measured by	Broca	Chantre	Broca	Chantre	Chantre
Number of Skulls	18	20	18	20	25
L	184	184	182	182	180
B	136	135	136	136	137
H'	129	131	133	—	—
B'	93	95	91	97	94
J	130	129	126	130	127
DC	29	25	26	23	23
O_1'	37	38	38	40	39
O_2	33	33	31	33	32
NH'	51	52	49	52	49
NB	24	26	25	26	25
fml	35	34	34	34	34
fmb	29	29	29	29	30
U	—	—	510	—	—
$100\,B/L$	73·9	73·4	74·7	74·7	76·1
$100\,H'/L$	70·1	71·2	73·1	—	—
$100\,B/H'$	105·4	103·1	102·3	—	—
$100\,(B-H')/L$	3·8	2·2	1·6	—	—
$100\,O_2/O_1'$	89·2	86·8	81·6	82·5	81·1
$100\,NB/NH'$	47·1	50·0	51·0	50·0	51·0
$100\,fmb/fml$	82·9	85·3	85·3	85·3	88·2

* All the indices in this Table were calculated from the mean values of the component lengths.

footnote || to Table I, p. 14). Ignoring the shorter series, there are five available ranging from the 11th Dynasty to Ptolemaic times. The mean measurements given in Table XIII were taken from various tables in (15) and (16), p. 507. All the series appear to be intermediate between the Upper and Lower Egyptian types: the breadths and cephalic indices arrange them in orders which are almost identically the same as the chronological order. As far as we can tell all the series are pure Egyptian and their relationships seem to confirm the hypotheses that were deduced from more adequate data.

8. *Upper Egyptian Series later than the Fifth Dynasty measured by German Craniometricians.* These are of very considerable importance because the majority are long series and an adequate number of measurements is provided for all of them. The individual measurements of two of the most important have been furnished by Dr Emil Schmidt in the Leipzig portion of the German Anthropological Catalogue (19). The longer (*loc. cit.* pp. 57—117) was reduced by Professor Karl Pearson and Dr Alice Lee and the means were first quoted in Fawcett's paper ((6), p. 426). The skulls came from Thebes and belonged to the

TABLE XIV.

Mean Measurements of Middle and late Dynastic Series of Egyptian Crania. Males.

Date	Middle Dynasties	Early and Middle Dynasties	Modern	18th to 21st Dynasties	18th to 20th Dynasties	18th Dynasty	18th and 19th Dynasties	26th to 30th Dynasties
Locality	El-Kubanieh North	El-Kubanieh South	Abyssinian Tigre Dist.	Thebes	Thebes	Thebes?	Abydos	Gizeh
Measured by	Toldt	Toldt	S. Sergi	Schmidt	Stahr	Schultz and Oetteking	Schmidt	Davin and Pearson
C	1355·0 (27)	1366·1 (67)	—	1387·6 (164)	1451·0 (50)	1393·8 (21)	1413·7 (25)	1438·9 (753)
L	182·2 (37)	182·7 (70)	183·7 (65)	181·9 (169)	183·5 (57)	185·1 (25)	183·1 (33)	185·3 (895)
L'	181·8 (37)	182·5 (66)	—	—	—	—	—	184·6 (872)
B	134·2 (37)	135·9 (70)	136·3 (65)	136·6 (169)	137·3 (56)	137·6 (25)	139·2 (33)	138·9 (896)
B'	90·4 (37)	92·7 (70)	94·6 (65)	93·8 (168)	94·9 (59)	95·1 (25)	96·2 (33)	94·8 (935)
H	—	—	—	134·1 (169)	—	134·3 (24)	132·4 (28)	134·1 (884)
H'	135·9 (35)	132·9 (65)	133·5 (64)	131·9 (168)	133·7 (54)	133·9 (25)	130·9 (33)	[132·4 (884)]
OH	116·1 (36)	115·0 (66)	112·6 (65)	114·3 (169)	114·1 (56)	113·8 (25)	113·5 (33)	113·4 (887)
LB	101·6 (34)	100·1 (63)	99·8 (64)	100·6 (169)	101·9 (56)	101·8 (25)	100·7 (33)	101·6 (896)
Q'	—	—	309·8 (65)	306·1 (169)	—	307·8 (25)	308·1 (32)	307·2 (878)
S	372·7 (33)	371·8 (69)	372·7 (58)	372·4 (168)	372·1 (51)	374·8 (25)	374·5 (32)	371·9 (884)
S_1	126·0 (36)	128·2 (69)	130·2 (65)	—	125·7 (54)	128·5 (25)	—	127·4 (928)
S_2	130·9 (33)	127·8 (67)	128·9 (65)	—	129·8 (50)	129·6 (25)	—	128·5 (908)
S_3	115·1 (30)	116·0 (66)	113·7 (57)	—	114·7 (54)	117·8 (25)		115·2 (889)
S_3'	97·3 (30)	97·8 (66)	95·8 (57)	96·5 (166)	96·1 (54)	97·8 (25)	96·3 (31)	94·7 (885)
U	505·6 (35)	506·2 (69)	513·0 (65)	510·8 (168)	512·6 (54)	512·8 (25)	517·1 (33)	518·7 (876)
$G'H$	70·1 (29)	70·2 (62)	69·1 (65)	68·8 (169)	70·1 (56)	69·7 (25)	71·3 (30)	70·4 (845)
GB	94·9 (29)	95·2 (57)	93·1 (57)	95·7 (161)	95·5 (54)	95·3 (24)	95·9 (25)	95·2 (877)
J	125·5 (26)	125·7 (54)	125·0 (42)	128·3 (166)	128·6 (53)	130·7 (25)	131·3 (33)	128·7 (785)
NH	—	—	—	50·61 (169)	—	51·1 (25)	52·1 (31)	51·65 (898)
NH'	49·8 (37)	49·4 (65)	48·9 (64)	[50·4 (169)]	50·8 (56)	[50·9 (25)]	[51·9 (31)]	[51·5 (898)]
NB	24·8 (36)	25·0 (63)	25·0 (63)	25·26 (166)	25·6 (53)	25·1 (25)	24·9 (26)	24·36 (893)
DC	23·4 (34)	23·3 (59)	22·0 (65)	—	21·1 (52)	24·3 (11)	24·2 (31)	—
SC	—	—	9·53 (64)	—	10·8 (52)	—	—	—
O_1	40·8 (36)	41·3 (66)	40·0 (65)	—	—	—	—	41·13 (886)
O_1'	38·1 (35)	38·4 (64)	[37·3 (65)]	38·14 (169)	38·3 (59)	38·3 (18)	38·7 (30)	[38·4 (886)]
O_2	32·3 (35)	32·8 (66)	33·2 (65)	32·96 (169)	32·8 (53)	33·0 (24)	33·5 (31)	33·73 (888)
G_1'	46·4 (25)	47·0 (55)	—	46·25 (160)	46·2 (43)	47·5 (23)	46·2 (26)	46·28 (821)
G_2	—	—	—	40·83 (151)	—	41·6 (21)	41·1 (21)	38·81 (703)
GL	96·5 (25)	98·1 (55)	95·2 (64)	96·41 (167)	96·1 (53)	96·0 (25)	96·2 (31)	95·03 (832)
fml	34·9 (32)	34·7 (64)	36·1 (57)	35·10 (168)	35·2 (55)	36·4 (25)	35·9 (31)	35·56 (894)
fmb	29·2 (32)	28·9 (64)	29·5 (59)	29·80 (165)	29·7 (54)	30·7 (25)	30·2 (30)	29·96 (905)
$100\,B/L'$	73·7 (37)	74·5 (66)	—	—	—	—	—	75·34 (866)
$100\,H/L$	—	—	—	{73·7 (169)}	—	{72·6 (24)}	73·7 (33)	72·46 (871)
$100\,B/L$	73·7 (37)	74·4 (70)	{74·2 (65)}	75·07 (169)	{74·8 (56)}	74·4 (25)	76·0 (33)	75·06 (884)
$100\,B/H$	—	—	—	{101·9 (169)}	—	{102·5 (24)}	103·1 (33)	103·70 (873)
$100\,H'/L$	74·7 (35)	72·9 (65)	{72·7 (64)}	{72·5 (168)}	{72·9 (54)}	72·3 (25)	71·5 (33)	{71·5 (884)}
$100\,B/H'$	{98·7 (35)}	{102·3 (65)}	{102·1 (64)}	{103·6 (168)}	{102·7 (54)}	102·8 (25)	106·4 (33)	{104·9 (884)}
$100\,(B-H')/L$	{−0·9 (35)}	{1·6 (65)}	{1·5 (64)}	{2·6 (168)}	{2·0 (54)}	{1·8 (24)}	{4·5 (33)}	{3·5 (884)}
$100\,G'H/GB$	{73·9 (29)}	{73·7 (57)}	{74·2 (57)}	72·32 (155)	{73·4 (54)}	{73·1 (24)}	{74·4 (25)}	74·13 (823)
$100\,NB/NH$	—	—	—	50·07 (166)	—	{49·1 (25)}	47·9 (30)	47·31 (881)
$100\,NB/NH'$	49·9 (36)	50·8 (63)	{51·1 (63)}	{50·1 (166)}	{50·4 (53)}	{49·3 (25)}	{48·0 (26)}	{47·4 (893)}
$100\,O_2/O_1$	{79·2 (35)}	{79·4 (66)}	{83·0 (65)}	—	—	—	—	82·12 (878)
$100\,O_2/O_1'$	84·9 (35)	85·7 (64)	{89·0 (65)}	86·5 (169)	{85·6 (53)}	{86·2 (18)}	86·8 (30)	{87·8 (886)}
$100\,fmb/fml$	84·0 (31)	83·4 (63)	{81·7 (57)}	85·50 (165)	{84·4 (54)}	84·4 (25)	84·3 (28)	84·47 (889)
$100\,G_2/G_1'$	—	—	—	88·5 (151)	—	87·6 (19)	89·3 (20)	{83·9 (703)}
$P\angle$	85°·6 (28)	84°·9 (58)	—	85°·5 (47)	85°·2 (55)	85°·9 (23)	—	85°·8 (795)
$N\angle$	{65°·2 (25)}	{67°·5 (55)}	{65°·4 (64)}	{66°·1 (167)}	{64°·5 (53)}	{65°·0 (25)}	{65°·2 (30)}	63°·9 (818)
$A\angle$	{73°·4 (25)}	{71°·0 (55)}	{73°·2 (64)}	{73°·0 (167)}	{74°·0 (53)}	{74°·0 (25)}	{72°·2 (30)}	74°·1 (818)
$B\angle$	{41°·4 (25)}	{41°·5 (55)}	{41°·4 (64)}	{40°·9 (167)}	{41°·5 (53)}	{41°·0 (25)}	{42°·5 (30)}	41°·9 (818)
θ_1	{29°·2 (25)}	{27°·6 (55)}	—	{28°·4 (47)}	{30°·3 (53)}	{28°·8 (23)}	—	30°·03 (795)
θ_2	{12°·2 (25)}	{13°·9 (55)}	—	{12°·5 (47)}	{11°·2 (53)}	{11°·9 (23)}	—	11°·89 (795)
Oc. I.	{62·3 (30)}	{62·0 (66)}	{61·9 (57)}	—	{61·4 (54)}	{60·6 (25)}	—	61·54 (884)
W (gms.)	—	—	605·2 (65)	—	598·4 (48)	—	—	—
EOW	—	—	—	—	—	97·4 (25)	—	—

Notes to Table XIV.

The indices and angles in curled brackets were calculated from the mean values of the component lengths. The lengths H', O_1' and NH' in squared brackets were found from the similar measurements H, O_1R and NH, R on the assumption that the differences would be the same as for Motley's very similar First Dynasty series (in Table V, p. 24, above). Those differences are very similar to values found for other ancient Egyptian series and, as far as we can tell, they are quite constant for all modern races of man. The capacities given by Stahr were probably determined by a method which would make them not comparable with the others. Tildesley (*Biometrika*, Vol. xiii. p. 217) has shown that in taking auricular heights on the Ranke's craniophor with the flute-topped plugs which are at present used in the Biometric Laboratory[*], a correction should be made by subtracting 2·5 mm. from the scale readings, and that is confirmed by comparison of mean direct and type contour measurements. That correction has been made to Motley's First Dynasty means in Table V. In the case of the series measured by Sergi and Toldt it is doubtful whether the nasal height measurement corresponds to our NH or NH': the difference between the two is small. Schmidt took the length to the lowest point of the pyriform aperture, i.e. whichever is greater of NH, R and NH, L, so the mean would be very slightly greater than that of either NH, R or NH, L. Oetteking does not say on which side he took the measurement. The Occipital Index is a complex function of the arc S_3 and the chord S_3', so it is not surprising that its mean value calculated from the indices of individual skulls should not be the same as the index found from the mean values of S_3 and S_3'. For all races the former exceeds the latter and the differences are 0·5 for the 26th to 30th Dynasties series, 1·0 for the Naqada A and Q series, 0·9 for the Naqada B, T and R series and 1·1 for Motley's First Dynasty series. It can be shown that the difference is not dependent on the number of skulls in the series so we may accept the mean of these four as a corrective factor for Egyptian types. The Occipital Indices in curled brackets in Table XIV were obtained by adding 0·9 to the indices deduced from the mean values of S_3 and S_3'.

In finding the means of the Abyssinian series, four juvenile skulls (Nos. 7, 13, 53 and 59) were excluded and one juvenile (No. 15) was excluded from Toldt's El-Kubanieh South series. The GL length of Schmidt's 18th to 21st Dynasties series was given as 95·79 (169); omitting one juvenile skull (No. 630) I find that it is 96·41 (167) and the alteration affects the angles of the fundamental triangle.

Several mean measurements of the 26th to 30th Dynasties series other than those in Table XIV were determined: viz. $F = 182·8$ (895), $Q = 307·5$ (876), $O_1L = 40·75$ (880), $O_2L = 33·83$ (888) (the O_1, O_2 and $100\,O_2/O_1$ in the table refer to the right orbit), $G_1 = 49·70$ (817), $100\,H/L' = 72·72$ (863), $100\,O_2/O_1$, $L = 83·06$ (876), $100\,G_2/G_1 = 78·36$ (668).

18th to 21st Dynasties. The means of the other series ((19), pp. 64—74), which belonged to the 18th and 19th Dynasties and came from Abydos, are now given for the first time. Schmidt gives quite complete measurements, following in general the *Frankfurter Verständigung*, but a few of them might have been defined with greater precision. The means are given in Table XIV. Both series are predominantly of the Lower Egyptian type and they show very close affinities with some previously considered.

When comparing the Coefficients of Racial Likeness calculated for several series of different lengths it is of the utmost importance to take into consideration the number of skulls on which each series is based. If that is done the five series of Lower Egyptian type shown in the upper part of Table XV are arranged by their inter-relations in the order shown. Schmidt's Abydos type stands very close to those of the 26th to 30th Dynasties and of the Royal Tombs of the First and Second Dynasties, but his longer series from Thebes of the same date is far closer to the Roman Period skulls. The former is of quite pure Lower Egyptian type and the latter diverges slightly from that race in the direction of the Upper Egyptian. The latter point is brought out clearly by a comparison with the five types shown in the lower half of Table XV, from the 18th Dynasty Shekh Ali to the Early

[* Earlier measures were taken with the *pointed* plugs for which the craniophor is scaled, and I believe there is not adequate ground for modifying their values as Tildesley proposed.—Ed.]

TABLE XV.

Coefficients of Racial Likeness with Schmidt's Series.

	18th and 19th Dynasties Abydos, Schmidt (29·9)				18th to 21st Dynasties Thebes, Schmidt (162·4)			
	All Characters		Indices and Angles		All Characters		Indices and Angles	
	C. L.	Number of Characters	C. L.	Number of Characters	C. L.	Number of Characters	C. L.	Number of Characters
26th to 30th Dynasties (885·4)	2·01 ± ·09	27	2·80 ± ·16	9	20·55 ± ·09	28	24·57 ± ·15	10
18th and 19th Dynasties Schmidt (29·9)	—	—	—	—	2·49 ± ·09	29	2·41 ± ·15	10
1st and 2nd Dynasties Royal Tombs* (21·8)	0·81 ± ·13	14	1·32 ± ·19	6	1·37 ± ·13	14	0·27 ± ·19	6
Roman Periods* (49·3) ...	1·74 ± ·13	14	1·05 ± ·19	6	1·23 ± ·13	14	0·11 ± ·19	6
18th to 21st Dynasties Schmidt (162·4)	2·49 ± ·09	29	2·41 ± ·15	10	—	—	—	—
1st Dynasty Royal Tombs Motley (33·2)	2·77 ± ·09	29	2·21 ± ·15	10	7·48 ± ·09	30	4·50 ± ·14	11
4th and 5th Dynasties* (39·9)	2·89 ± ·13	14	2·53 ± ·19	6	4·70 ± ·13	14	1·77 ± ·19	6
18th Dynasty Shekh Ali* (40 0)	7·43 ± ·13	14	7·45 ± ·19	6	4·28 ± ·13	14	5·38 ± ·19	6
6th to 12th Dynasties* (168·6)	15·73 ± ·13	14	19·03 ± ·19	6	15·84 ± ·13	14	19·16 ± ·19	6
1st Dynasty Private Tombs (33·6)*	8·90 ± ·13	14	14·98 ± ·13	6	6·77 ± ·13	14	9·80 ± ·19	6
Late Pre-dynastic* (88·5)	13·66 ± ·13	14	19·41 ± ·13	6	14·07 ± ·13	14	22·05 ± ·19	6
Early Pre-dynastic* (35·8)	16·91 ± ·13	14	21·81 ± ·19	6	13·67 ± ·13	14	19·55 ± ·19	6

* These series were measured by Thomson and MacIver.

Pre-dynastic. They represent the earlier transitional stages from the pure Upper Egyptian type to the distinct one of Lower Egypt. The Coefficients with Schmidt's series are all of a higher order than the ones in the upper part of the table, but the order of arrangement is in both cases the same as that which was arrived at in considering the inter-relations of the earlier series themselves. It is quite reasonable to place the 6th to 12th Dynasties series before that of the 1st Dynasty Private Tombs since the disparity between the numbers of skulls in each is so great. Remembering that Schmidt's 18th to 21st Dynasties series is far longer than his other, we may conclude that its relationships with the Upper Egyptian types, though remote, are distinctly closer than are those between the latter and Schmidt's Middle Dynastic series from Abydos. A detailed comparison of individual measurements need hardly be made; it would confirm the conclusions that have already been reached. The greatest values of α are almost invariably found between the characters $100\,B/H'$, B, $100\,B/L$ and L, while H', $N\angle$ and $100\,NB/NH'$ show differences that are smaller but still significant in some cases; and certain other characters are reasonably constant throughout. Nevertheless Schmidt's L and $G'H$ of the 18th to 21st Dynasties series are as small as for any other dynastic Egyptian series and their values differ significantly from those of some closely related types. This alone would be very meagre evidence of alien admixture, but it will be shown later that there are better reasons for supposing that the series in question had diverged slightly from the pure Egyptian stock.

In 1919 C. Toldt (20) furnished individual measurements of Egyptian skulls that had been disinterred at El-Kubanieh near Assuan. The site is 100 miles from Thebes and is more southerly than any other with which we are at present concerned. The whole material came from two graveyards and it was divided into the following groups (see (20), pp. 526, 635 and 636):

El-Kubanieh South	Early Dynastic	62 skulls.
	Middle Dynastic	52 skulls.
El-Kubanieh North	Middle Dynastic	31 skulls.
	Nomads	26 skulls.

The nomads in the second group were supposed to be related to tribes of the northern Sudan and they apparently belonged to Middle Dynastic times. Unfortunately it is not possible to separate them from the other skulls from the El-Kubanieh North cemetery. The means of the two groups reduced from Toldt's individual measurements are given in Table XIV. The Coefficients between them and the most closely related series are in Table XVI.

TABLE XVI.

Coefficients of Racial Likeness with Toldt's Series from El-Kubanieh.*

		6th to 12th Dynasties (168·6)	El-Kubanieh North (33·1)	12th to 15th Dynasties (65·9)	El-Kubanieh South (63·5)	18th Dynasty Shekh Ali (40·0)
El-Kubanieh South (63·5)	14 Characters ... 6 Indices and Angles	5·90 ± ·13 7·78 ± ·19	3·24 ± ·13 6·33 ± ·19	2·95 ± ·13 4·61 ± ·19	— —	2·86 ± ·13 5·30 ± ·19
El-Kubanieh North (33·1)	14 Characters ... 6 Indices and Angles	0·81 ± ·13 −0·14 ± ·19	— —	1·41 ± ·13 2·60 ± ·19	3·24 ± ·13 6·33 ± ·19	1·69 ± ·13 2·67 ± ·19

* All the Coefficients in Table XVI are calculated for the 14 characters, of which 6 are indices and angles, available for all Thomson and MacIver's series. Between the two El-Kubanieh series 28 characters can be used giving a Coefficient of 1·60 ± ·09 and of 3·16 ± ·14 for the 11 indices and angles.

It is clear that the El-Kubanieh North and South series are very similar to the Middle Dynastic people of Upper Egypt: they take up the positions shown in Table XVI and Fig. 1. The former is the later in time and yet it has been less changed from the Upper Egyptian type of the First Dynasty than the other and the difference is doubtless to be attributed to the nomadic element. In calculating the Coefficients in Table XVI the greatest values of α were found for the characters $100\,B/H'$, B, $100\,B/L$, H', $N\angle$, $100\,H'/L$ and NH'. The prominence of the last two is rather exceptional. The El-Kubanieh North series, but not the other, has an unusually large basio-bregmatic height and height index, but other characters are not distinctive from other Middle Dynastic types. The Coefficients between Toldt's two series and Schmidt's Theban skulls of the 18th to 21st Dynasties are given in Table XVII.

TABLE XVII.

Coefficients of Racial Likeness between Schmidt's 18th to 21st Dynasties and Toldt's El-Kubanieh Series.

	18th to 21st Dynasties Schmidt (162·4)			
	All Characters		Indices and Angles	
	C. L.	Number of Characters	C. L.	Number of Characters
El-Kubanieh North (33·1) ...	4·50 ± ·09	27	6·14 ± ·15	10
El-Kubanieh South (63·5) ...	2·65 ± ·09*	27	3·27 ± ·15*	10
18th Dynasty Shekh Ali (40·0) ...	4·28 ± ·13	14	5·38 ± ·19	6

* When these Coefficients are computed for the 14 characters, of which 6 are indices and angles, available for Thomson and MacIver's series, they are increased to 8·51 ± ·13 and 4·23 ± ·09 respectively.

That the El-Kubanieh South series should be closer to Schmidt's 18th to 21st Dynasties than the El-Kubanieh North is quite consistent with the relationships already found, but that it should also be closer than the 18th Dynasty series from Shekh Ali is unexpected. This discordance might be attributed to the fact that Toldt's series was heterogeneous in respect to time, but it will be shown later that a more probable explanation is that both it and Schmidt's 18th to 21st Dynasties series were intermixed very slightly with an alien stock. The positions shown in Fig. 1, p. 2, of the El-Kubanieh relative to the other series are the most reasonable.

We regret that it has only been possible to make slight use of Dr Bruno Oetteking's paper (22) on ancient Egyptian crania. He has not sexed his material and therefore it cannot be used for comparative purposes. It is strange to find a craniological study as late as 1909 in which the writer is content to take means and variations on a population without distinction of sex. The Egyptian skulls measured by Dr Oetteking are in the University of Zürich and in 1918 Dr Adolf Schultz published measurements of 35 of them (25 ♂ and 10 ♀) for the purpose of a special enquiry ((23), measurements on pp. 72—79). Accepting that sexing, the means of the 25 specimens were reduced from Schultz and Oetteking's tables and they are given in Table XIV. The material was probably Theban of the 18th Dynasty. The Coefficients of Racial Likeness are given in Table XVIII.

The short Zürich series is of almost identically the same type as Thomson and MacIver's of the First and Second Dynasties. It takes up a position between the latter and Schmidt's series of the 18th and 19th Dynasties (see Fig. 1, p. 2) and it gives another representation of the wide-spread population of Lower Egyptian type which was predominating in Upper Egypt in Late Dynastic times. There is not the slightest suggestion of alien admixture.

TABLE XVIII.

Coefficients of Racial Likeness with Schultz and Oetteking's Zürich Series.

	Zürich Series (23·9)			
	All Characters		Indices and Angles	
	C. L.	Number of Characters	C. L.	Number of Characters
18th and 19th Dynasties Schmidt (29·9)	0·51 ± ·09	29	1·25 ± ·15	10
1st and 2nd Dynasties Royal Tombs (21·8)	0·13 ± ·13	14	0·25 ± ·19	6
26th to 30th Dynasties (885·4)	2·02 ± ·09	31	1·84 ± ·14	12

In 1907 Dr Hermann Stahr (21) published full individual measurements of 113 skulls from Thebes belonging to the period from the 18th to 20th Dynasties. The means of the 56 skulls given in Table XIV were reduced from Stahr's tables; not a single one is given in the original paper. The series appears to be quite a homogeneous one and it is of almost pure Lower Egyptian type. The Coefficients of Racial Likeness between it and the series which it most closely resembles are given in Table XIX.

TABLE XIX.

Coefficients of Racial Likeness with Stahr's 18*th to* 20*th Dynasties Series.*

	18th to 20th Dynasties Stahr (54·1)			
	All Characters		Indices and Angles	
	C. L.	Number of Characters	C. L.	Number of Characters
Roman Period (49·3)	1·42 ± ·13	14	0·62 ± ·19	6
1st and 2nd Dynasties Royal Tombs (21·8)	0·91 ± ·13	14	1·44 ± ·19	6
18th Dynasty Schultz and Oetteking (23·9)	− 0·06 ± ·09	28	− 0·54 ± ·14	11

The bond between Stahr's series and that measured by Schultz and Oetteking is closer than that between any other two series dealt with in this paper. That is not surprising as they were contemporaneous and both came from Thebes.

For the series dealt with in the Biometric Laboratory and those measured by German craniometricians many more characters are available than the 21 provided by Thomson and MacIver. The significances of the differences of the additional characters that are used in the computation of the Coefficients of Racial Likeness can be gauged in the usual way by considering the values of α. No very detailed comparison need be made here: it entirely confirms the conclusions already arrived at. In the following list (Table XX) the types are arranged in order from the almost pure Upper Egyptian type represented by the Naqada crania to the pure Lower Egyptian type of the 26th to 30th Dynasties series. The intermediate series, all of which came from Upper Egypt, represent the more or less gradual transition from one extreme to the other. Where a character does not differ significantly between the Upper and Lower Egyptian types it is always found that no one of the intervening types differs significantly from either of the extremes. Where, however, the characters of the Upper Egyptian type do differ significantly from those of the Lower then, for those characters, there is a more or less gradual transition from one extreme to the other when the series are arranged in the order shown, which is the order in which the Coefficients of Racial Likeness arrange them. For some characters Motley's First Dynasty series is quite clearly differentiated from the others.

The length characters S, S_1, S_2, S_3, S_3', G_1' and *fml*, the facial breadths GB and the foraminal indices, profile angles and occipital indices of the Upper and Lower Egyptian and all intermediate types seem to be identically the same. Small differences are found between the characters associated with facial heights (O_2, $100 O_2/O_1'$ and $100 G'H/GH$) and greater differences between those influenced by change in calvarial breadth (B', U, Q' and C), and the gradual transition from the Upper to the Lower Egyptian type can be traced more or less clearly in the case of the latter. In general a comparison of these additional characters entirely confirms the relationships of the two extreme types already noted. The Lower Egyptian race is characterised by its much greater calvarial and zygomatic breadths and cephalic index and its slightly greater facial heights, while all lengths in the sagittal plane and facial breadths other than the zygomatic breadth are the same for the two races. Some mean measurements of the First Dynasty skulls of the Royal Tombs from Abydos are given in Table XX. It has been shown that that type was primarily Lower Egyptian, though inclining slightly towards the Upper

TABLE XX.

Comparison of Additional Mean Measurements.

Series	Constant Characters				Characters showing significant differences					
	S	G_1'	GB	$O_1'R$	B'	U	O_2R	100 $G'H/GB$	100 O_2/O_1'	C
Naqada *A* and *Q* Series ...	372·6	48·0	95·9	39·1	92·6	510·4	32·2	72·7	82·4	1388·6
El-Kubanieh North	372·7	46·4	94·9	38·1	90·4	505·6	32·3	73·9	84·9	1355·0
El-Kubanieh South	371·8	47·0	95·2	38·4	92·7	506·2	32·8	73·7	85·7	1366·1
18th to 21st Dynasties Schmidt	372·4	46·3	95·7	38·1	93·8	510·8	33·0	72·3	86·5	1387·6
18th to 20th Dynasties Stahr ...	372·1	46·2	95·5	38·3	94·9	512·6	32·8	73·4	85·6	1451·0
18th Dynasty Schultz and Oetteking	374·8	47·5	95·3	38·3	95·1	512·8	33·0	73·1	86·2	1393·8
18th and 19th Dynasties Schmidt	374·5	46·2	95·9	38·7	96·1	517·1	33·5	74·4	86·8	1413·7
26th to 30th Dynasties ...	371·9	46·3	95·2	38·4	94·8	518·7	33·7	74·1	87·8	1438·9
1st Dynasty Royal Tombs Motley	369·0	49·5	97·3	40·0	93·0	513·5	33·5	76·8	86·1	1364·3
Modern Abyssinian, Tigre district	372·7	—	93·1	—	94·6	513·0	33·2	74·2	—	—

Egyptian type. The majority of its measurements, including all those of the calvaria, are intermediate between the measurements of the Upper and Lower Egyptian races, but others clearly differentiate the First Dynasty skulls from all other Egyptian types. The greater facial heights $G'H$ and NH', but not O_2, are the most distinctive characters, but $B\angle$, GB, $O_1'R$, G_1' and G_2 are also greater than for any other series and the indices $100\,G'H/GB$ and $100\,fmb/fml$ diverge from the normal type. Motley's series is undoubtedly not pure Egyptian, but we are at present unable to say what other race took part in its make up; it was certainly not negroid.

In the last row of Table XX there are some mean measurements of modern Abyssinian skulls; the complete measurements are in Column 4 of Table XIV. The data are provided by Dr Sergio Sergi (26). The series with which we are concerned came from the extreme north of the country—the Tigre district—and it is distinctly different from two other shorter ones from Central Abyssinia and Somali respectively*. There is a close relationship between the Tigre skulls and Middle Dynastic Egyptian types.

When it is remembered that an interval of well over 3000 years must have separated the Abyssinians from any of the dynastic skulls with which they are compared in Table XXI, the Coefficients appear to be surprisingly low. The bond with the El-Kubanieh South series is at least as close as that between the Moorfields and Whitechapel English series of 17th century Londoners†. At the same time we should have anticipated even lower Coefficients if the Abyssinians

TABLE XXI.

Coefficients of Racial Likeness with Modern Abyssinians (Tigre District).

		18th to 21st Dynasties Schmidt (162·4)	Ptolemaic Period (76·6)	18th Dynasty Shekh Ali (40·0)	El-Kubanieh South (63·5)	12th to 15th Dynasties (65·9)
Modern Abyssinian (Tigre District) (61·8)	14 Characters ...	3·48 ± ·13	6·02 ± ·13	2·10 ± ·13	1·16 ± ·13	3·72 ± ·13
	6 Indices and Angles	1·76 ± ·19	2·01 ± ·19	2·49 ± ·19	3·27 ± ·19	3·24 ± ·19
	23 Characters ...	3·57 ± ·10	—	—	1·87 ± ·10	—
	8 Indices and Angles	4·12 ± ·17	—	—	2·46 ± ·17	—

---

* There is not the slightest suggestion of negroid admixture in the series from Central Abyssinia and Somali.

† These series are in *Biometrika*, Vol. v. p. 92. The Coefficients between the Whitechapel (87·2) and Moorfields (24·2) English are 2·05 ± ·09 for 26 characters and 1·73 ± ·16 for 9 indices and angles. We have good reason to suppose that the Coefficients between the Abyssinian and El-Kubanieh South series should be even lower than the values shown. The greatest differences between them are for $N\angle$ and $A\angle$ and for these characters the modern series is similar to all middle dynastic ones and it is very probable that the aberrant values given by Toldt are due to a difference in method of measurement. If $N\angle$ and $A\angle$ are excluded the two series have Coefficients of 0·95 ± ·10 for 21 characters and - 0·13 ± ·19 for 6 indices and angles.

were of pure ancient Egyptian type. In Table XXI the series are arranged in the order given by their own inter-relation but their degrees of resemblance to the Abyssinians do not arrange them in the same order. The Coefficients with Schmidt's 18th to 21st Dynasties series are decidedly lower than we should have anticipated. It is probable that both that series and the El-Kubanieh South were slightly mixed with an alien stock which had a greater effect on the later Abyssinians. The only individual characters which distinguish the latter from the normal middle dynastic type of Upper Egypt are its small *NH′*, *G′H*, *LB* and *GB*, and the nasal index being peculiarly large and the foraminal index peculiarly small. And for those characters the El-Kubanieh South series and Schmidt's of the 18th to 21st Dynasties show a tendency to diverge from the normal in the direction of the Abyssinian values: the *NH′* and 100*fmb*/*fml* of the former and the *G′H* of the latter are smaller than for any other dynastic series*.

9. *Type Contours of First Dynasty Skulls from Royal Tombs at Abydos.* The horizontal, transverse and sagittal type contours of the First Dynasty skulls from royal tombs at Abydos (see p. 22 above) were constructed by Mr G. H. Motley from the mean measurements of contours of the individual skulls in the manner commonly practised in the Biometric Laboratory†. The mean measurements are given in Tables XXII, XXIII and XXIV and the outlines in Figs. 3—8. The ♂ contours may be compared with the types of 100 ♂ skulls from Lower Egypt which formed part of the long 26th to 30th Dynasties series (see p. 29 above) and of which the three type contours were constructed by Benington in 1911 (*Biometrika*, Vol. VIII.

TABLE XXII.

Mean Measurements of the Transverse Vertical Contours of 32 ♂ *and* 10 ♀ *First Dynasty Skulls from Royal Tombs at Abydos.*

Sex	*MA*	1*R* = 1*L*	*M* ¼*R*	*M* ¼*L*	2*R*	2*L*	3*R*	3*L*	4*R*	4*L*	5*R*	5*L*	6*R*	6*L*
♂	110·6	57·1	60·0	59·3	60·5	60·1	62·9	62·7	65·3	64·6	66·1	64·9	66·0	64·7
♀	108·2	54·3	57·1	56·5	58·1	57·7	60·9	60·4*	63·1	62·4*	64·3	62·9	64·4	62·9

Sex	7*R*	7*L*	8*R*	8*L*	9*R*	9*L*	10*R*	10*L*	*A* ¼*R*	*A* ¼*L*	*ZR . R*		*ZR . L*	
											y	*x*	*y*	*x*
♂	65·1	63·3	61·6	59·2	54·0	50·5	39·4	36·0	19·2	17·8	60·2	3·1	59·8	3·7
♀	63·0	60·9	59·7	56·4	52·0	48·0	37·5	34·4	18·9	17·6	57·4	2·6	57·3	3·1

* These lengths are the means of 9 contours.

---

* The Abyssinian and the El-Kubanieh South series are the only two with which we are dealing to have a greater frontal (S_1) than parietal arc (S_2): see Table XIV. These measurements are unfortunately not available for Schmidt's series.

† For a full account of the method used see *Biometrika*, Vol. XIV. pp. 227—241.

pp. 173—177). The sagittal contour of the latter series was only given for the arc above the nasio-gamma base line, so there is no representation of the facial bones and base of the skull. On superposing the outlines, the differences between them are seen to be very small: the type zones would actually overlap for the whole of their lengths, so we cannot say that any of the differences are significant. This conclusion is in accordance with that derived from a comparison of direct measurements as the only significant differences which could be distinguished in that way were those between facial measurements.

TABLE XXIII.

Mean Values of the Horizontal Contours of 32 ♂ *and* 10 ♀ *First Dynasty Skulls from Royal Tombs at Abydos.*

Sex	*FO*	*F ¼R*	*F ¼L*	*F ½R*	*F ½L*	2*R*	2*L*	2 ½*R*	2 ½*L*	3*R*	3*L*	4*R*	4*L*	5*R*	5*L*
♂	183·9	23·0	23·5*	34·7	35·5†	45·9	47·2†	46·1	46·8*	47·9	48·0†	53·7	53·8§	61·0	60·8*
♀	177·5	20·7	21·9	32·9	34·6	44·3	45·8	44·1	45·0	46·0	46·6	51·9‡	51·9	58·9	58·8

Sex	6*R*	6*L*	7*R*	7*L*	8*R*	8*L*	9*R*	9*L*	10*R*	10*L*	*O ¼R*	*O ¼L*	*T(R)x*	*T(L)x*	*T(R)y*	*T(L)y*
♂	65·8	65·2*	66·9	66·6*	64·3	64·0	57·7	57·8	45·1	45·8	24·4	24·9	20·5	19·9†	47·9	48·9†
♀	63·6	63·9	65·6‡	65·8	63·4‡	64·3	56·9	58·4	45·9	46·4	25·5	25·0	19·9	19·7	45·9	47·2

* These lengths are the means of 31 contours.
† ,, ,, ,, 30 ,,
‡ ,, ,, ,, 9 ,,
§ ,, ,, ,, 29 ,,

TABLE XXIV.

Mean Measurements of the Sagittal Contours of 31 ♂ *and* 10 ♀ *First Dynasty Skulls from Royal Tombs at Abydos.*

Sex	$N\gamma$	Ordinates above $N\gamma$													
		$O=N$	$N\frac{1}{4}$	1	2	3	4	5	6	7	8	9	$\gamma\frac{1}{4}$	$\gamma\frac{1}{8}$	$\gamma=10$
♂	182·8	15·4	31·9	56·3	69·3	76·7	80·8	82·4	83·0	80·0	70·4	51·4	25·8	18·8	5·27
♀	176·5	15·2	36·2	56·5	68·4	74·6	78·3	80·2	80·0	76·3	67·3	46·2	23·3	16·2	-1·52*

* The negative sign here indicates that the 10th ordinate meets the outline below the $N\gamma$ line on the type.

TABLE XXIV—(*continued*).

Sex	Ordinates below $N\gamma$								Vertex		β		Glabella	
	$O=N$	$N\frac{1}{4}$	1	2	8	9	$\gamma\frac{1}{2}$	$\gamma\frac{1}{4}$	x from N	y	x from N	y	x from N	y
♂	70·7*	67·2	57·4	54·5†	46·1	35·1	26·3	19·0	103·8	83·4	75·0	81·1	1·2	9·7
♀	65·9	61·5	52·6	49·7	48·3	38·4	30·4	23·3	95·7	81·1	75·3	79·0	1·0	9·1

Sex	λ		Sub-Orb.		Aur.		Op.		Inion		Basion		Alv.	
	x from γ	y	x from N	y	x from γ	y	x from γ	y	x from γ	y	from γ	from N	from N	from Bas.
♂	5·7	27·7	10·9	29·1	94·1	28·1	57·0*	53·3*	14·3	32·1	106·6	101·5	74·4	95·7
♀	4·8	22·2	9·8	28·7	90·9	27·9	56·0	55·1	14·3	35·5	102·5‡	98·7‡	70·1‡	93·4‡

Sex	Nose			Palate				Sp.		N. S.		Max. Sub. of Alv. N. S. Chord	
				P'		P							
	(i)	(ii)	(iii)	x from N	y	x from Alv.	y	x from N	y	x from N	y	x from N	y
♂	1·02†	3·0†	6·3*	49·8†	51·5†	36·1†	19·6†	67·0	34·0	6·4	54·1	3·4	62·5
♀	0·82	2·9	6·7	48·0	48·0	38·2	18·6	64·6‡	32·8‡	6·0	49·8	4·0	54·9

Sex	Frontal Max. Sub. to $N\beta$		Occipital Max. Sub. to λ Op.		Max. Sub. to $N\lambda$		Max. Sub. to G. I.	
	x from N	y	x from λ	y	x from N	y	x from G	y
♂	50·9	24·6	45·2§	26·8§	84·9	68·7	104·5	100·0
♀	50·2	26·7	42·5	26·2	81·9	69·2	100·0	98·4

* These lengths are the means of 29 contours.

† ,, ,, ,, 30 ,,

‡ ,, ,, ,, 9 ,,

§ ,, ,, ,, 28 ,,

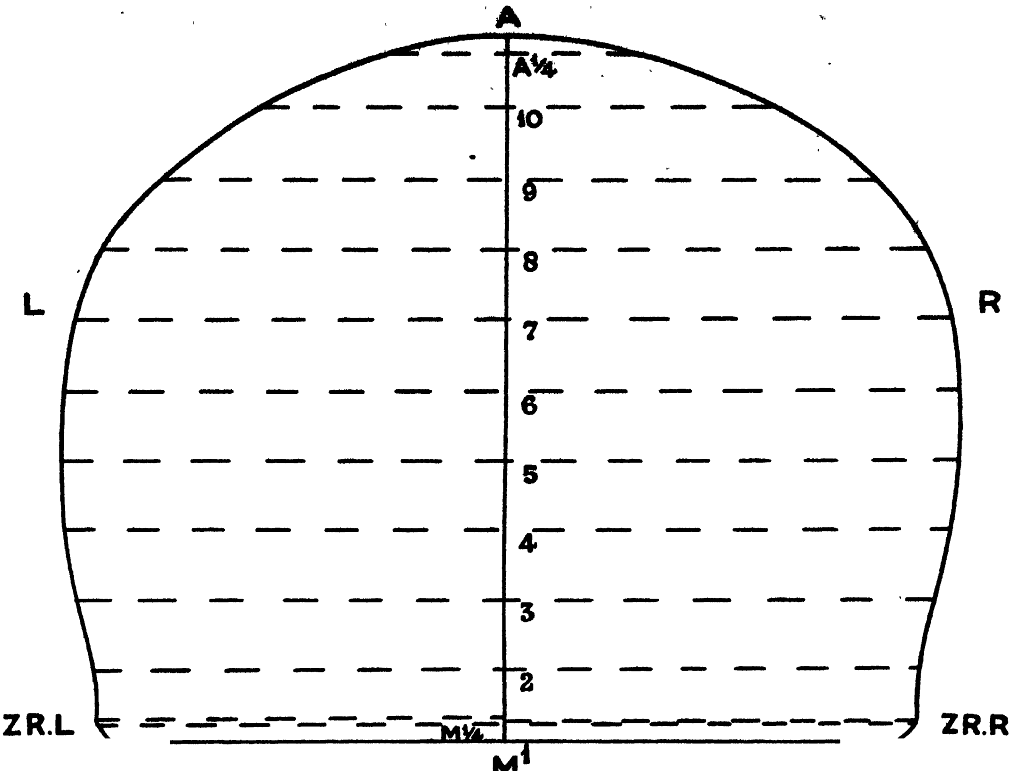


Fig. 3. Transverse Type Contour of 32 ♂ First Dynasty Skulls from Abydos.

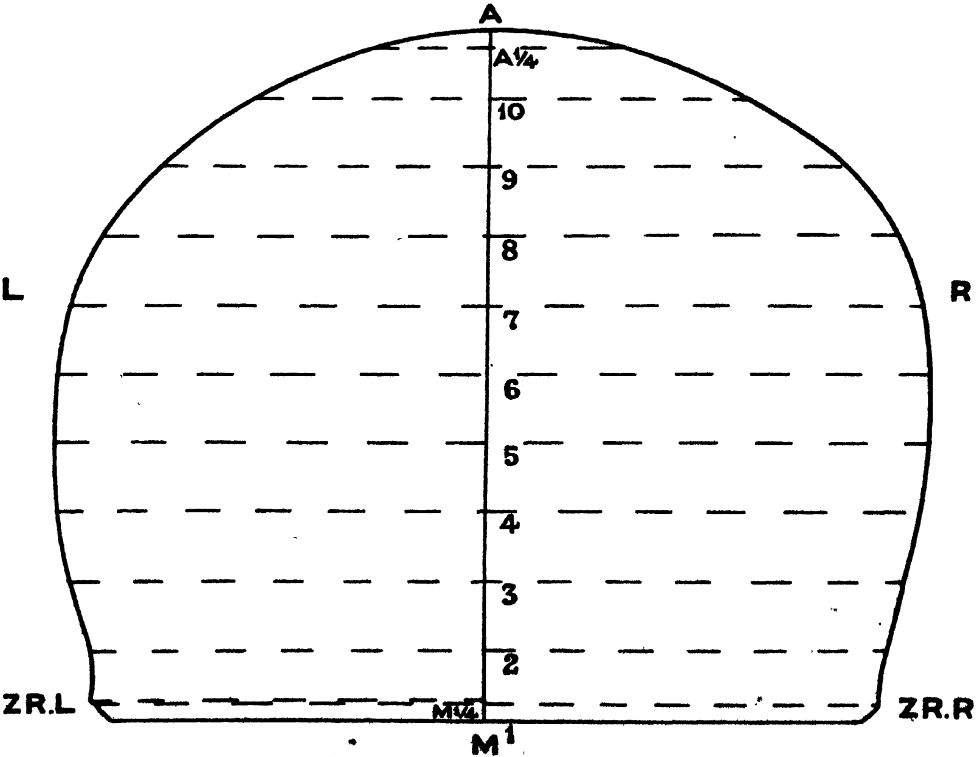


Fig. 4. Transverse Type Contour of 10 ♀ First Dynasty Skulls from Abydos.

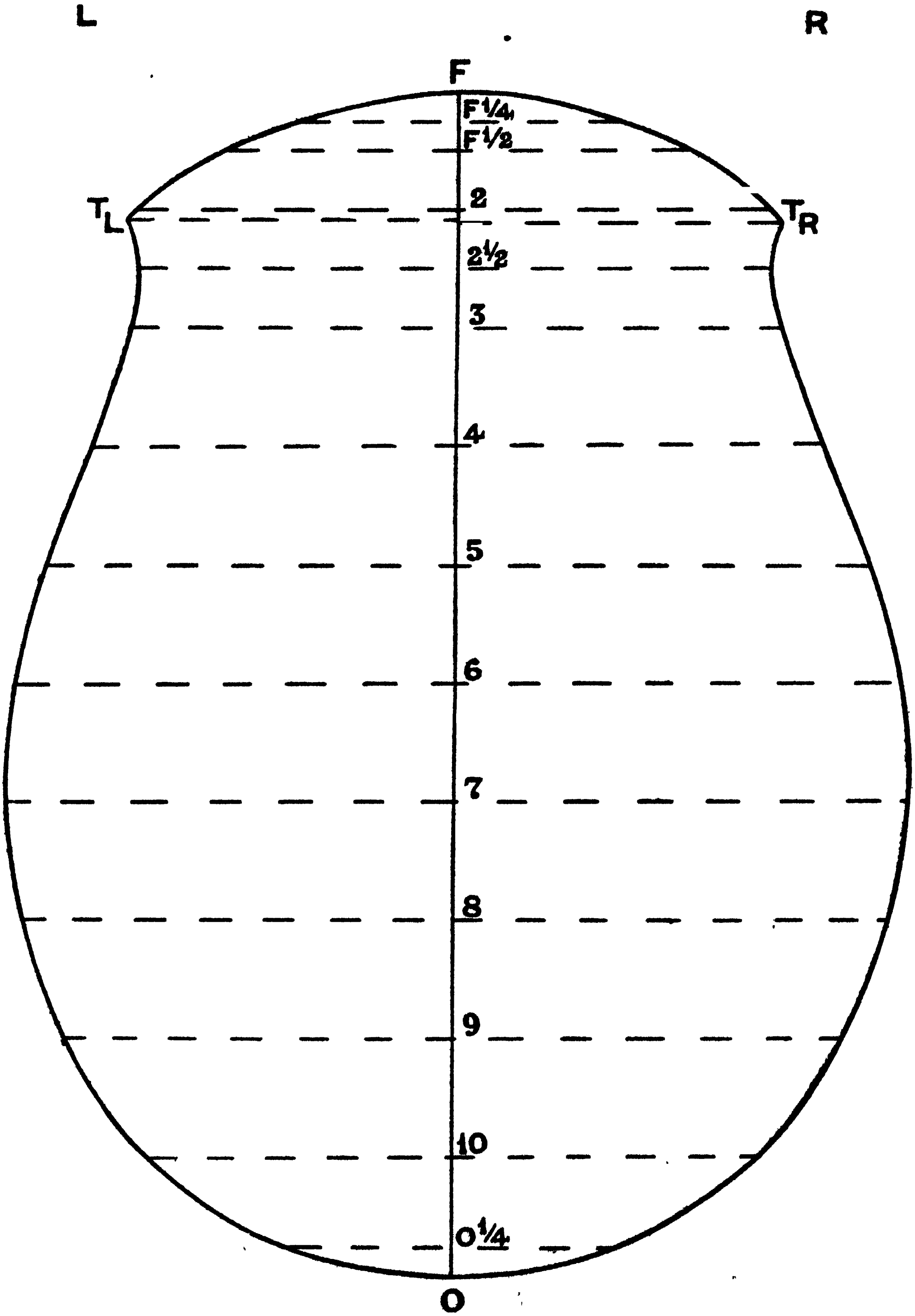


Fig. 5. Horizontal Type Contour of 32 ♂ First Dynasty Skulls from Abydos.

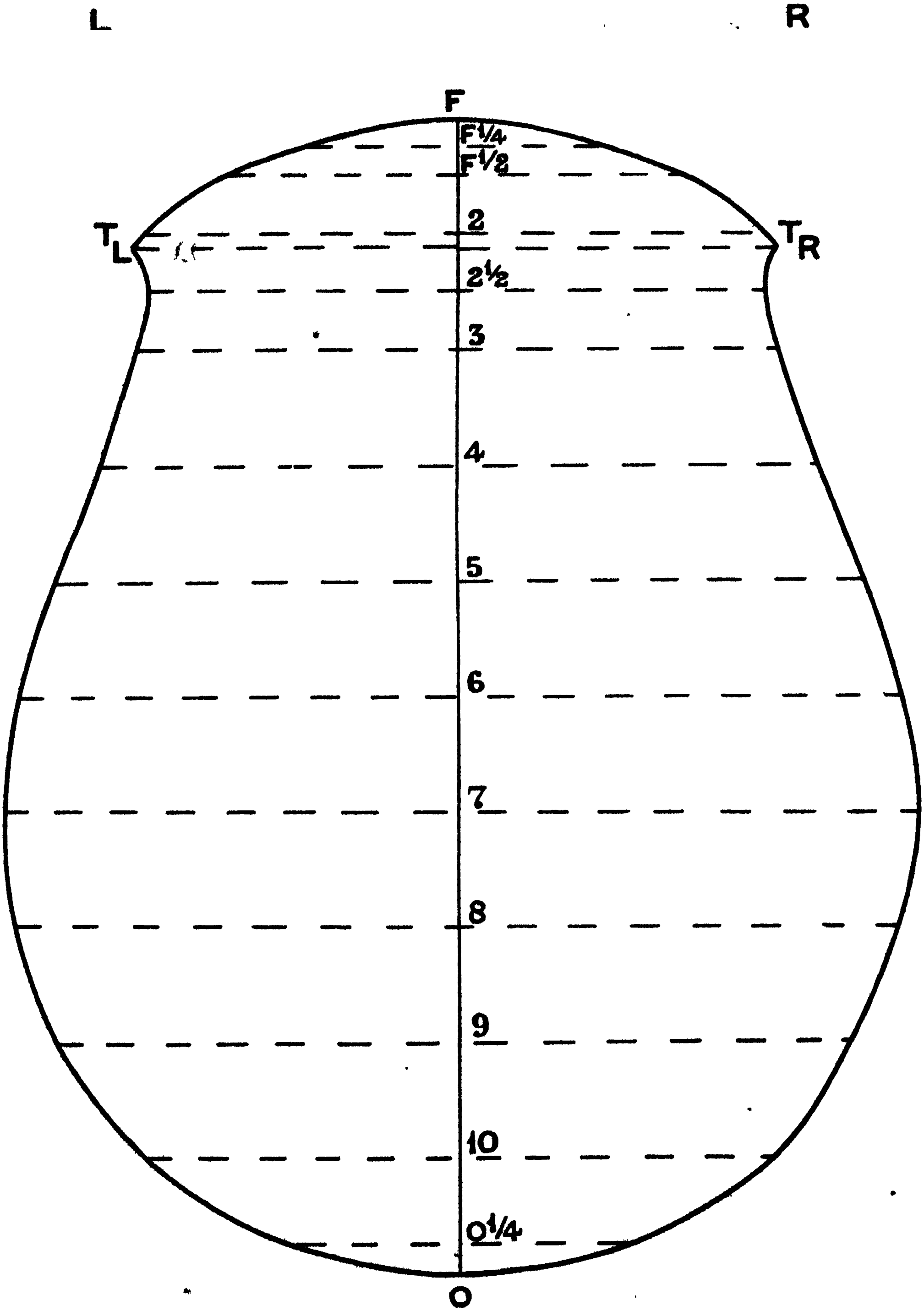


Fig. 6. Horizontal Type Contour of 10 ♀ First Dynasty Skulls from Abydos.

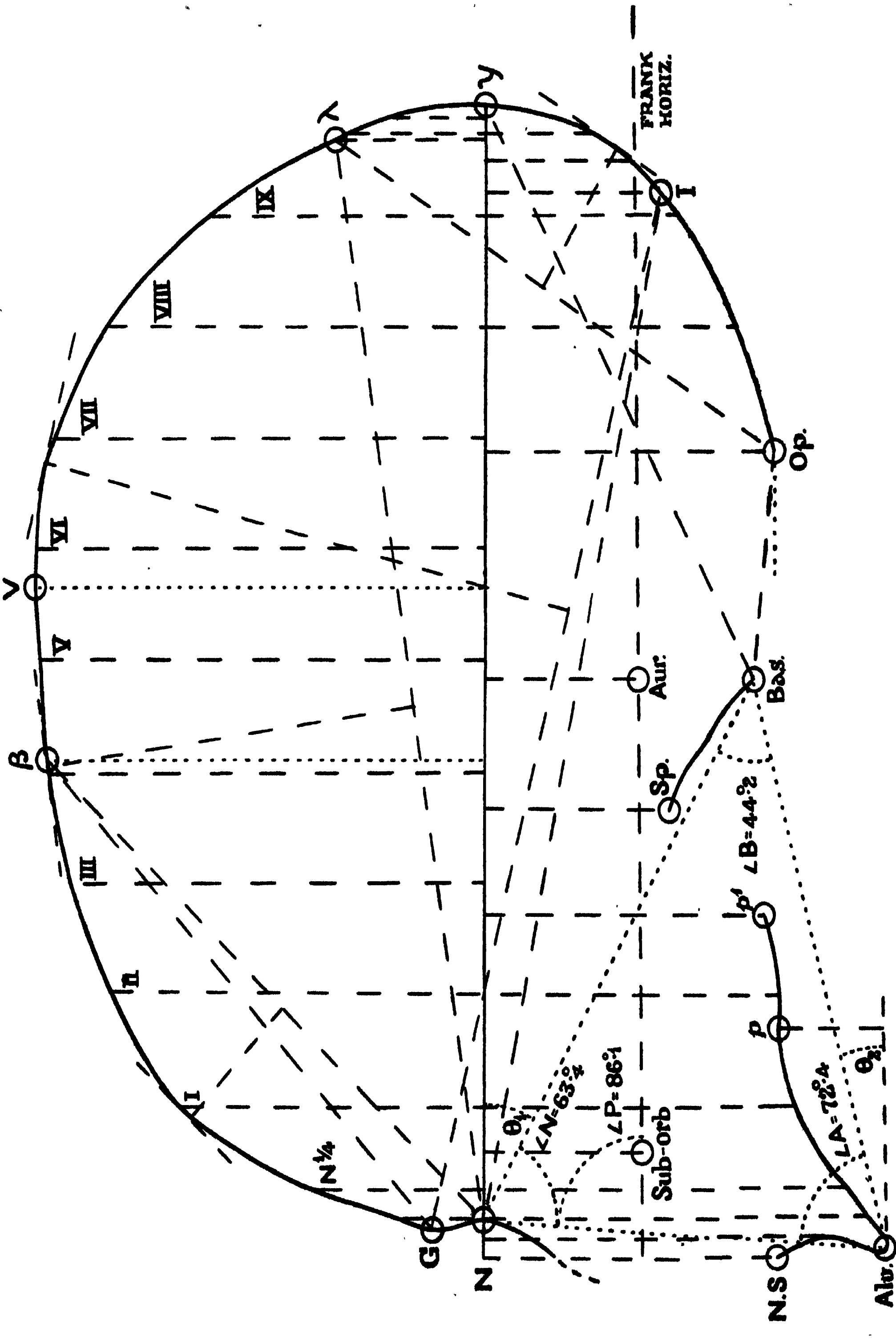


Fig. 7. Sagittal Type Contour of 32 ♂ First Dynasty Skulls from Abydos.

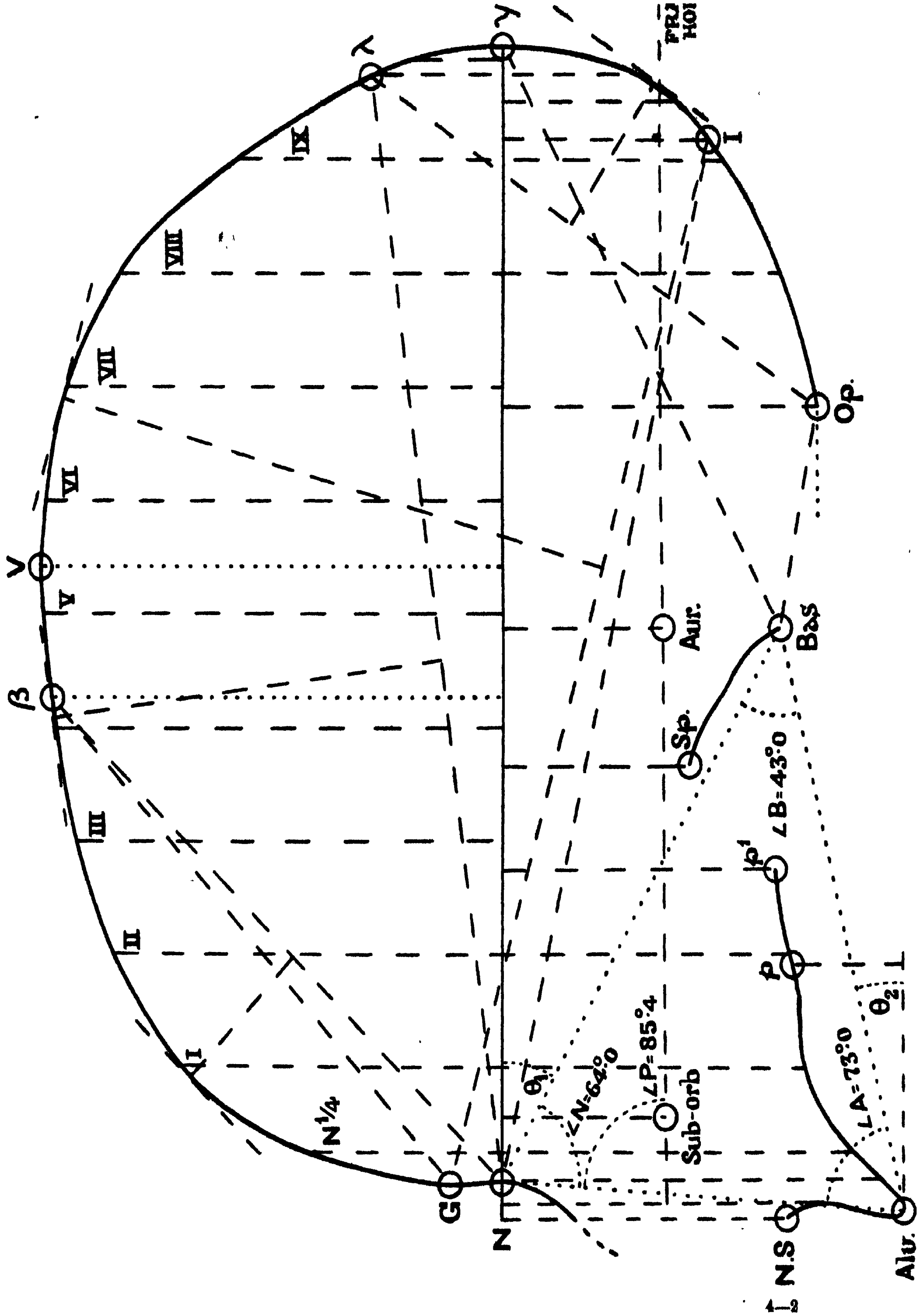


Fig. 8. Sagittal Type Contour of 10 ♀ First Dynasty Skulls from Abydos.

REFERENCES.

(1) Morton, S. G. Observations on Egyptian Ethnology, derived from Anatomy, History, and the Monuments. *Transactions of the American Philosophical Society*, Vol. IX. Philadelphia. 1846. (Paper read in 1842 and 1843.)

(2) Newberry, P. E. *Egypt as a Field for Anthropological Research.* Presidential Address to Section H (Anthropology). *Report of the British Association for the Advancement of Science*, 1923.

(3) Flinders Petrie, W. M. (with Quibell, J. E.). *Naqada and Ballas.* 1896.

(4) —— *Diospolis Parva.* "The Cemeteries of Abadiyeh and Hu." 1901.

(5) —— *Gizeh and Rifeh.* 1907.

(6) Fawcett, C. D. A Second Study of the Variation and Correlation of the Human Skull with special reference to the Naqada Crania. *Biometrika*, Vol. I. pp. 408—467. 1902.

(7) Crewdson Benington, R. and Pearson, Karl. A Study of the Negro Skull with Special Reference to the Congo and Gaboon Crania. *Ibid.* Vol. VIII. pp. 292—339. 1912.

(8) Smith, H. D. A Study of Pygmy Crania, based on skulls found in Egypt. *Ibid.* Vol. VIII. pp. 262—266. 1912.

(9) Warren, E. An Investigation on the Variability of the Human Skeleton: with special reference to the Naqada Race. *Phil. Trans.* Series B, Vol. 189, pp. 135—227. 1897.

(10) Morgan, J. de. *Recherches sur les Origines de l'Égypte.* "L'Age de la Pierre et les Métaux." 1896.

(11) —— *Recherches sur les Origines de l'Égypte.* "Ethnographie Préhistorique et Tombeau Royal de Négadah." 1897.

(12) Randall-MacIver, D. *The Earliest Inhabitants of Abydos.* 1901.

(13) Thomson, A. and Randall-MacIver, D. *The Ancient Races of the Thebaid.* 1905.

(14) Chantre, E. Étude Craniologique sur la population prépharaonique de la Haute Égypte. *Bulletin de la Société d'Anthropologie de Lyon.* Tome 18, pp. 187—192. 1899.

(15) —— *Recherches Anthropologiques dans l'Afrique Orientale.* "Égypte." 1904.

(16) Quatrefages, A. de and Hamy, E. T. *Crania Ethnica.* "Les Cranes des Races Humaines." 1882.

(17) Pruner-Bey. Recherches sur l'Origine de l'ancienne Race Égyptienne. *Mémoires de la Société d'Anthropologie de Paris.* T. 1, pp. 399—433. 1861—3. (Paper read in 1861.)

(18) Biasutti, R. "Crania Aegyptiaca." Esame di 42 Crani di Egiziani Antichi. *Archivio per l'Antropologia e la Etnologia.* Vol. 35, pp. 323—362. Firenze. 1905.

(19) Die Anthropologischen Sammlungen Deutschlands. (Dr Emil Schmidt's catalogue of his own collection at Leipzig was published with *Archiv für Anthropologie.* Bd. XVII. 1887.)

(20) Toldt, C. Anthropologische Untersuchung der Menschlichen Ueberreste aus den Altägyptischen Gräberfeldern von El-Kubanieh. *Denkschriften der Akademie der Wissenschaften in Wien.* Mathematisch-naturwissenschaftliche Klasse. Bd. 96, pp. 593—672. 1919.

(21) Stahr, Hermann. *Die Rassenfrage in Antiken Aegypten.* "Kraniologische Untersuchungen an Mumienköpfen aus Theben." Leipzig. 1907.

(22) Oetteking, B. Kraniologische Studien an Altägypten. *Archiv für Anthropologie.* Bd. XXXVI. pp. 1—90. 1909.

(23) Schultz, A. Anthropologische Untersuchungen an der Schädelbasis. *Ibid.* Bd. XLIV. pp. 1—103. 1918.

(24) Widenmann. Untersuchung von 30 Dschaggaschädeln. *Ibid.* Bd. XXV. pp. 361—396. 1898.

(25) Drontschilow, Krum. Metrische Studien an 93 Schädeln aus Kamerun. *Ibid.* Bd. XL. pp. 161—183. 1913.

(26) Sergi, Sergio. *Crania Habessinica. Contributo all' Antropologia dell' Africa Orientale.* 1912.

NOTE.—*The Definition of the Alveolar Point* (see p. 20 above). By KARL PEARSON.

There has been much vagueness about the localisation of the so-called alveolar point. It culminates probably in Flower's definition[*]: "*Alveolar Point*, the centre of the anterior margin of the upper alveolar arch." This might appear capable of interpretation had not Flower defined what we term the profile length and then described the *Basialveolar Length* as "Basion to alveolar point, or the most distant part of the anterior margin of the alveolar arch."

Now the most distant part of the anterior margin is rarely in the mesial plane of the skull, but on the anterior face of one or other of the front incisors on either side of the mesial plane. This difficulty has been recognised by certain authorities, and they define the alveolar point to be the point in which a tangent to the border of the alveolar arch on these incisors meets the mesial plane. This is a fairly definite point, but as it usually "hangs in the air" it is not capable of accurate determination in the case of the average skull; it is not at all a suitable definition for the *Knotenpunkt* of a whole series of cranial measurements. We next turn to the classical French craniometricians. Paul Broca in his *Instructions craniologiques et craniométriques*, p. 42, Paris 1875[†], writes as follows:

> La *région incisive*, comprise entre le nez et la bouche, est formée par la partie de l'arcade alvéolaire qui supporte les dents incisives. On y remarque, sur la ligne médiane, la *suture incisive*, dont l'extrémité supérieure aboutit à l'épine nasale et dont l'extrémité inférieure, correspondant au milieu de l'arcade alvéolaire, se nomme le *point alvéolaire*.

The alveolar point is here defined to be the "lower extremity" of the *incisive suture*, and this is said to correspond to the middle of the "arcade alvéolaire." But where does the incisive suture terminate? Some might hold that it runs as far as the incisive foramen, but no one has ventured to place the alveolar point there. At any rate if it be an *extremity* of the suture we are seeking, it is more likely to be on the posterior than the anterior face of the alveolar process.

Topinard defines the *Point alvéolaire* as the "point le plus antérieur et le plus déclive du bord alvéolaire supérieur[‡]."

Unfortunately in the great bulk of crania the two characteristics by which Topinard defines the alveolar point are directly contradictory, the most anterior point is not the lowest point.

Turning to English writers we have already referred to Flower. Cunningham[§] identifies the Alveolar Point with the Prosthion and defines it as "the centre of the anterior margin of the upper alveolar margin." This leaves us in a state of perplexity as to how a margin can have a margin, much less two margins, as presumably if the upper alveolar margin has an anterior margin it has also a posterior margin. Should the definition be translated "the point where the anterior margin of the upper alveolar process cuts the mesial plane," we must point out that the alveolar margin between any pair of adjacent teeth is extraordinarily vague and often lacks determination, and, according to the personal equation of the observer, may lie anywhere from the geodesic linking up the mid anterior points of the alveolar margin on the incisors to the lowest point of the central process between the incisors. Quain's *Anatomy* is more concise; it reads*:

---

[*] *Catalogue of the R.C. of S. Museum, Homo Sapiens*, Part I, pp. xiii and xviii, 1879.

[†] See also *Mémoires de la Société d'Anthropologie*, T. II. 2e série, pp. 96 *et seq.*

[‡] *Éléments d'Anthropologie Générale*, p. 251, 1885.

[§] *Text-book of Anatomy*, p. xvi, Appendix D, 1909.

" Below the nasial aperture is the *intermaxillary suture*; the lower point of this, on the alveolar margin between the central incisors, is named the *alveolar point* (*prosthion*)."

This definition is more or less a return to Broca's where we have reference to the incisive suture and its lower point (if not its " extremity"); the lower point is said to be on the alveolar margin, but then this margin does not terminate the suture and the position of the margin between the central† incisors is the matter under discussion.

Turning to the Germans we naturally consult first the *Frankfurter Verständigung* (*Archiv für Anthropologie*, Bd. XV. 1884), but the only clue to the alveolar point is to be found in the definition of the upper face height (*Obergesichtshöhe*). This we take from nasion to alveolar point. The *Verständigung* defines it as " von der Mitte der Sutura naso-frontalis, bis zur Mitte des Alveolarrandes des Oberkiefers zwischen den mittleren Schneidezähnen " (S. 3). This certainly seems identical with the lowest point of the central process between the middle incisors. The diagrams attached to the *Verständigung* throw no more light on the *accurate* localisation of the alveolar point than those in Broca, Topinard, Cunningham or Quain.

E. Schmidt in his *Anthropologische Methoden*, 1885, gives S. 210 the following definition: "Alveolarpunkt (Point alvéolaire), Kreuzungspunkt des Zahnrandes des Oberkiefers mit der Medianebene." Here again we ask what is the " Zahnrand," the tooth-border, at the mesial plane? R. Martin in his *Lehrbuch der Anthropologie* (S. 514, 1914) sees something of the difficulty and separates the Prosthion from the point to which the upper face height is to be taken. Thus: " Prosthion [pr] = derjenige Punkt am Alveolarrand des Oberkiefers, der in der Mediansagittal-Ebene zwischen den mittleren Schneidezähnen am meisten nach vorn vorragt. Das Prosthion liegt also nicht an dem unteren Ende des zwischen die Schneidezähne vorgeschobenen Knochenteils, sondern an der am meisten nach vorn vorstehenden Stelle seiner Vorderwand. Nur für die Messung der Obergesichtshöhe ist der Messpunkt an die Spitze des genannten Fortsatzes zu legen."

Thus Martin agrees with our interpretation of the *Frankfurter Verständigung*. But he proceeds at once to contradict himself for he defines (S. 553) the *Obergesichtshöhe* (or, the naso-alveolar height for he cites the English term) as the " Geradlinige Entfernung des Nasion vom Prosthion," thus introducing confusion into the *Verständigung*, as well as into his own book. But is it always the case that the most forward projecting point of the alveolar border lies between the middle incisors? We think not. Again, is that border generally distinguishable when it meets the mesial plane? Martin appears to identify it with the most projecting point of the anterior wall of the central process between the middle incisors; but it is the exact boundary between the alveolar arch and the central process between the mid incisors which is so difficult of determination.

---

* *Elements of Anatomy*, Vol. IV. Part I, p. 118, 1915.

† We have heard a stringent critic assert that there cannot be *two* central or *two* mid incisors, which is perfectly correct, but the term is convenient, and what is good enough for Quain is permissible for us.

But this is not the greatest difficulty. If there be two points used, the prosthion and the alveolar point—defined, as the Biometricians have defined it since 1895, as the lowest point of the central process between the middle incisor teeth—then the computation of the angles and sides of the fundamental triangle of the skull becomes impossible. Martin would measure the upper face height (according to the *Frankfurter Verständigung*) to the alveolar point, but the profile length to the prosthion[*], and he would take his profile angle from the nasion-prosthion line[†]. It is clear that there could be no orientation of the fundamental triangle (*Gesichtsdreieck*) unless all the measurements were taken to the same point, and accordingly somewhere between his pages 514 and 562, Martin has given up the alveolar point as the extremity of his upper face height. The *Verständigung* itself is very obscure on the measurement of the Upper Face Height and of the Profile Length. We have already given its definition of the former. The latter length is thus defined (S. 4): "von dem am meisten vorspringenden Punkt der Mitte des äusseren Alveolarrandes[‡] des Oberkiefers bis zum vorderen Rand des Foramen magnum (in der Medianebene) gemessen." It is by no means clear that the alveolar extremities of the upper face height and the profile length are to be taken to the same point; it looks as if the former was to be taken to the alveolar point and the latter to a point which may or may not be the prosthion.

It was experience of the indefiniteness and positive confusion in this matter and the need for selecting a point from which *all* measurements may be taken, and from which the orientation of the fundamental triangle (*Gesichtsdreieck*) with regard to the Frankfurt horizontal plane may be determined, which led to our selection of the mid-point of the central process between the middle incisors as the alveolar point. As a matter of fact we have preferred to define it not as the mid-point, but as the *lowest* point of the central process when the skull is orientated to the Frankfurt plane[§]. As far as the fundamental triangle is concerned there is no more valid reason for selecting the anterior than the posterior face of the alveolar arch. I regret only that we did not take palate lengths from this alveolar point rather than from a point on the posterior alveolar margin as vague as the prosthion on the anterior alveolar margin.

---

[*] *Loc. cit.* S. 550.

[†] *Loc. cit.* S. 560.

[‡] There seems here to be some confusion of the alveolar border and alveolar arch; for if *Alveolarrand* signifies the alveolar border or lower margin of the alveolar arch, this is presumably a curve which can have only one middle point, *i.e.* intersects the mesial plane in a single point. What is therefore the sense of introducing the words "the most projecting point"? We can't choose a point within a point. It would appear therefore either that border is used for arch, and the point in question may lie anywhere on the incisive suture where we reach the greatest distance from the basion, or that the *Alveolarrand* is an area in the neighbourhood of the alveolar border on which area we are to choose the point most remote from the basion. How this area is separated from the anterior face of the alveolar arch we do not know, nor have we found it anywhere stated. Another alternative, namely that the mid-point of the alveolar margin is invariably the point furthest removed from the basion, would not only render the definition tautological, but it involves a dogma so obviously erroneous that it need not be discussed.

[§] The importance of this will be readily realised by the reader who has examined the central process in numerous crania; central processes occur which are wedge-shaped V rather than semi-circular ∪.

I realise that objection may be taken to measuring the profile angle as we have done, probably on the ground that the nasion-prosthion line is nearer to the profile slope in the living than is the nasion-alveolar point line. But both are so far off the living profile—owing to the flesh thickness of the lip—that the small horizontal projected distance from prosthion to alveolar point makes little difference compared with the great divergence between the cranial and living profiles. This argument, indeed, might justify the use of Cuvier's facial line to the tip of the incisors[*] as better in its turn than using the nasion-prosthion line.

If further justification be needed for my choice in 1895 of the alveolar point as on the central process of the middle incisors I think it may be found in the *Entente Internationale* of 1906, resulting from the Monaco Congress[†], for the unification of craniometric measures. Internationally this concordat should replace the *Frankfurter Verständigung*. Therein we read under the definition of the *Diamètre naso-alvéolaire*, our upper face height, that the lower extremity of this measurement is "le point le plus inférieur du bord[‡] alvéolaire entre les deux incisives médianes et supérieures."

Again for the *Diamètre alvéolo-basilaire*, our profile length, the anterior extremity is the "point alvéolaire, ou point médian du bord antérieur de l'arcade alvéolaire." Although the Frankfurt Concordat had been on trial for more than twenty years the distinguished craniometricians of the Committee of the Monaco Congress do not appear to have agreed on a common alveolar point for all their measurements or, if they did, they describe it in one of their definitions as the lowest point, and in the other as simply the mid-point on the alveolar margin, which itself has a very dubious existence in the mesial plane!

It is very clear that one and all these definitions contain fossilised fragments of earlier conceptions; they can hardly be the product of craniometricians, who have had to struggle with the actual difficulty of determining the prosthion or alveolar point on long series of crania, nor of those who realise the importance of eliminating personal equation.

We might define the prosthion as the point in which the geodesic[§] on the anterior face of the alveolar arch between the mid-points of the anterior faces of the middle incisors at the anterior alveolar border meets the incisive suture; and again the alveolar point as the point in which the incisive suture or the mesial plane meets the lowest element (relative to the Frankfurt horizontal) of the central process between the middle incisors. In other words it is the lowest point of the section of the above central process by the mesial plane. If the section be straight edged ⟍___⟋, then we must select its middle point as alveolar point.

We believe that our alveolar point admits of far readier determination than the prosthion, and when it cannot be ascertained owing to absorption of the alveolar borders, then also no prosthion is available.

---

[*] Or better still a tangent from the nasion to the alveolar arch in the mesial plane.

[†] *L'Anthropologie*, Tom. XVII. pp. 568—7.

[‡] Does "bord" here signify "arcade"? For if not there is again tautology if the "bord" is a curve meeting the mesial plane.

[§] In practice the geodesic is determined by drawing with a fine pencil the shortest line (curved not straight) between the mid-points of the anterior alveolar border on the middle incisors.

ON THE MOMENTS OF THE HYPERGEOMETRICAL SERIES.

By Professor V. ROMANOVSKY.

(University of Taskend.)

In a paper of the same title Prof. K. Pearson (*Biometrika*, Vol. XVI. pp. 157—162) has given a finite difference relation, which enables us to write down the successive moments of a hypergeometrical series about its mean. This relation is very elegant and leaves nothing to desire if we evaluate the moments one after another. But it does not enable us to write down any desired moment without knowing the preceding ones. The aim of this article is to provide a formula, which permits us to obtain the moments of a hypergeometrical series independently.

Let $\alpha = -r$, $\beta = -qn$, $\gamma = pn - r + 1$, and

$$F(\alpha, \beta, \gamma, x) = A\left[1 + \frac{r}{1}\,\frac{qn}{pn - r + 1}\,x + \frac{r(r-1)}{1\,.\,2}\,\frac{qn(qn-1)}{(pn-r+1)(pn-r+2)}\,x^2 + \dots\right],$$

where

$$A = \frac{pn(pn-1)\dots(pn-r+1)}{n(n-1)\dots(n-r+1)}.$$

Here the coefficient of x^t is the probability of drawing t individuals without some mark in a sample of size r drawn from a population of n containing pn marked and qn non-marked individuals.

Let $E(t)$ stand for the mathematical expectation of t*. Then

$$E(t) = rq,$$

and it is easy to see by successive differentiations of the function F, that

$$E[t(t-1)\dots(t-s+1)] = \left[\frac{d^sF}{dx^s}\right]_{x=1} \quad\dots\dots\dots\dots\dots\dots\dots(1),$$

or

$$U_s = \left[\frac{d^sF}{dx^s}\right]_{x=1} \quad\dots\dots\dots\dots\dots\dots\dots\dots\dots(1\ bis),$$

if we put

$$E[t(t-1)\dots(t-s+1)] = U_s \quad\dots\dots\dots\dots\dots\dots\dots(2).$$

The function $F(\alpha, \beta, \gamma, x)$ satisfies the well-known differential equation

$$x(1-x)\frac{d^2F}{dx^2} + (\gamma - \delta x)\frac{dF}{dx} - \alpha\beta F = 0,$$

where

$$\delta = \alpha + \beta + 1 = 1 - r - qn.$$

---

* If $F(-r, -qn, pn-r+1, x)$ be written $a_0 + a_1x + a_2x^2 + \dots + a_tx^t + \dots + a_rx^r$, we have $S_{t=0}^{t=r}(a_t) = 1$ and $E(f(t)) = S_{t=0}^{t=r}(a_tf(t))$, or $E(f(t))$ is the mean value of $f(t)$ ($t=0$ to r) weighted with the probability of drawing t unmarked individuals in the sample of r.

Differentiating this equation s times we have

$$x(1-x)\frac{d^{s+2}F}{dx^{s+2}}+s(1-2x)\frac{d^{s+1}F}{dx^{s+1}}-s(s-1)\frac{d^sF}{dx^s}+(\gamma-\delta x)\frac{d^{s+1}F}{dx^{s+1}}$$
$$-s\delta\frac{d^sF}{dx^s}-\alpha\beta\frac{d^sF}{dx^s}=0.$$

Put here $x=1$ and taking regard to (1) and (2), we have

$$(\gamma-\delta-s)\,U_{s+1}=[\alpha\beta+s\delta+s(s-1)]\,U_s,$$

or

$$(n-s)\,U_{s+1}=(r-s)(qn-s)\,U_s \text{ ....(3).}$$

Now $U_1=rq$ and thus, with the aid of the above relation, we find

$$U_s=\frac{r(r-1)\ldots(r-s+1)\,qn(qn-1)\ldots(qn-s+1)}{n(n-1)\ldots(n-s+1)} \text{ ....(4).}$$

The sth moment μ_s of our hypergeometrical series is defined by the relation

$$\mu_s=E(t-rq)^s.$$

This may be written

$$\mu_s=E[K_s(t)]$$
$$=E\left[K_s(0)+t\frac{\Delta K_s(0)}{1}+t(t-1)\frac{\Delta^2K_s(0)}{2!}+\ldots\right.$$
$$\left.+t(t-1)\ldots(t-s+1)\frac{\Delta^sK_s(0)}{s!}\right],$$

if we put

$$(t-rq)^s=K_s(t),$$

and apply the interpolation formula of Newton.

Therefrom

$$\mu_s=K_s(0)+\frac{\Delta K_s(0)}{1}U_1+\frac{\Delta^2K_s(0)}{2!}U_2+\ldots+\frac{\Delta^sK_s(0)}{s!}U_s \text{ ....(5).}$$

Here

$$\Delta^hK_s(0)=[\Delta^hK_s(t)]_{t=0}=\Delta^h0^s-C_s{}^1rq\Delta^h0^{s-1}+C_s{}^2(rq)^2\Delta^h0^{s-2}-\ldots \quad \text{...(6),}$$

as it is easy to see.

Let us put

$$\frac{\Delta^h0^\lambda}{h!}=h_\lambda \text{ ....(7).}$$

Applying the relation

$$\Delta^h0^\lambda=h^\lambda-C_h{}^1(h-1)^\lambda+C_h{}^2(h-2)^\lambda-\ldots+(-1)^{h-1}C_h{}^1\,.\,1^\lambda \text{ ;........(8),}$$

we easily deduce that

$$h_\lambda=h\,.\,h_{\lambda-1}+(h-1)_{\lambda-1} \text{ ....(9),}$$

for all positive integer values of h and λ. It must be remarked that

$$h_h=1 \text{ and } h_\lambda=0 \text{ for } h>\lambda \text{ ....(10).}$$

With the aid of these relations we easily find the values of h_λ for different values of h and λ. We place here a short table of these values.

Table of values of h_λ *for* $h, \lambda = 1, 2, \ldots, 10$.

h	λ									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	1
2	—	1	3	7	15	31	63	127	255	511
3	—	—	1	6	25	90	301	966	3025	9330
4	—	—	—	1	10	65	350	1701	7770	34105
5	—	—	—	—	1	15	140	1050	6951	42525
6	—	—	—	—	—	1	21	266	2646	22827
7	—	—	—	—	—	—	1	28	462	5880
8	—	—	—	—	—	—	—	1	36	750
9	—	—	—	—	—	—	—	—	1	45
10	—	—	—	—	—	—	—	—	—	1

Now put

$$\frac{\Delta^h K_s(0)}{h!} = L_{h,s} \qquad \text{.....................................(11).}$$

Then, according to (6) and (7),

$$L_{h,s} = h_s - C_s^1 h_{s-1} rq + C_s^2 h_{s-2} (rq)^2 - \ldots + (-1)^s C_s^s h_0 (rq)^s,$$

or, in consequence of (10),

$$L_{h,s} = h_s - C_s^1 h_{s-1} rq + C_s^2 h_{s-2} (rq)^2 - \ldots + (-1)^{s-h} C_s^{s-h} (rq)^{s-h} \quad \ldots(12).$$

The polynomials $L_{h,s}$ for $s \leqq 10$ can be written down by the aid of the table. For greater values of s we must recur to (8) or (9). Putting, for example, $s = 5$, we shall have

$$L_{0,5} = -(rq)^5,$$
$$L_{1,5} = 1 - 5rq + 10(rq)^2 - 10(rq)^3 + 5(rq)^4,$$
$$L_{2,5} = 15 - 35rq + 30(rq)^2 - 10(rq)^3,$$
$$L_{3,5} = 25 - 30rq + 10(rq)^2,$$
$$L_{4,5} = 10 - 5rq,$$
$$L_{5,5} = 1.$$

We see from (5) and (11) that

$$\mu_s = L_{0,s} + L_{1,s} U_1 + L_{2,s} U_2 + \ldots + L_{s,s} U_s \qquad \text{...............(13),}$$

which is the relation I proposed to establish.

For example, taking $s = 5$, we find

$$\begin{aligned} \mu_5 = {} & -(rq)^5 + [1 - 5rq + 10(rq)^2 - 10(rq)^3 + 5(rq)^4]\frac{rqn}{n} \\ & + [15 - 35rq + 30(rq)^2 - 10(rq)^3]\frac{r(r-1)\,qn(qn-1)}{n(n-1)} \\ & + [25 - 30rq + 10(rq)^2]\frac{r(r-1)(r-2)\,qn(qn-1)(qn-2)}{n(n-1)(n-2)} \\ & + [10 - 5rq]\frac{r(r-1)(r-2)(r-3)\,qn(qn-1)(qn-2)(qn-3)}{n(n-1)(n-2)(n-3)} \\ & + \frac{r(r-1)(r-2)(r-3)(r-4)\,qn(qn-1)(qn-2)(qn-3)(qn-4)}{n(n-1)(n-2)(n-3)(n-4)}. \end{aligned}$$

Putting in (13) $n = \infty$, we shall find the sth moment for a binomial $(p+q)^r$ about its mean

$$\mu_s = L_{0,s} + L_{1,s}\, rq + L_{2,s}\, r(r-1)\, q^2 + \ldots + L_{s,s}\, r(r-1)\ldots(r-s+1)\, q^s \ldots (14),$$

where the coefficients $L_{h,s}$ keep their values. This relation may be transformed into a more condensed form which is indicated in my note on the moments of a binomial $(p+q)^r$ (see *Biometrika*, Vol. XV. formula (5)).

It is not difficult to establish for hypergeometrical moments the relations

$$\lim\,(\mu_{2s}/\mu_2^{\,s}) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} x^{2s} e^{-x^2}\, dx,$$

$$\lim\,(\mu_{2s+1}/\mu_2^{\,s+\frac{1}{2}}) = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{+\infty} x^{2s+1} e^{-x^2}\, dx,$$

for $n \rightarrow \infty$ and $r \rightarrow \infty$. Therefore, according to the theorem of Tchebycheff-Markoff, we conclude that, for a hypergeometrical series, the probability of inequalities

$$t_1 \leqq \frac{t - rq}{\sqrt{\mu_2}} \leqq t_2,$$

for any given t_1 and t_2, tends to $\frac{1}{\sqrt{\pi}}\int_{t_1}^{t_2} e^{-x^2}\, dx$ for $n \rightarrow \infty$ and $r \rightarrow \infty$, i.e. for n and r sufficiently great the hypergeometrical distribution is near to the normal curve, as was long ago indicated by Laplace. (*Mémoires de mathématique...présentés à l'Académie*, Tom. VI. p. 6, 1774.)

[*Note* The two methods, i.e. deducing the higher moments in succession by Pearson's formula and deducing the higher moment directly by Romanovsky's formula, were applied to obtain the fifth moment when

$$n = 100,\ p = \tfrac{1}{10},\ q = \tfrac{9}{10} \text{ and } r = 10.$$

The first gave

$$\mu_2 = \tfrac{9}{10},\ \mu_3 = -\tfrac{288}{539},\ \mu_4 = 2{\cdot}156{,}743 \text{ and } \mu_5 = -4{\cdot}122{,}411 \text{ in succession.}$$

The second gave directly

$$\mu_5 = -4{\cdot}122{,}410.$$

There was very little difference in the time or labour required to reach μ_5, using an arithmometer. But the former gave with no greater expenditure of labour, μ_2, μ_3 and μ_4, which also would generally be required, and the work would be very much longer by the second method, if all the moments were required. If, however, Romanovsky's method were used to check the μ_5 of the first method it would at the same time check all the earlier moments, and this indicates a decided field of usefulness for the formula. ED.]

PROBABILITY INTEGRALS FOR A HYPERGEOMETRIC SERIES.

BY PROFESSOR BURTON H. CAMP, PH.D.

(1) IF a bag contains pn black balls and qn other balls, where $p+q=1$, the probabilities of obtaining $r, r-1, \ldots, 1, 0$ black balls in one drawing of r balls at a time are given by the terms of the finite hypergeometric* series

$$\frac{y^{(r)}}{n^{(r)}}\left\{1+\frac{r^{(1)}}{1!}\,\frac{n-y}{y-r+1}+\frac{r^{(2)}}{2!}\,\frac{(n-y)(n-y-1)}{(y-r+1)(y-r+2)}+\ldots\right\} \quad \text{.........(i)},$$

where $y=pn$, $n-y=qn$, $y^{(r)}=y(y-1)\ldots(y-r+1)$, etc. The probability of obtaining at least $s=r-t$ black balls is given by the sum of the first $(t+1)$ terms of (i), which may be written in the form

$$S_{t+1}(y)=\frac{1}{n^{(r)}}\left\{y^{(r)}+\frac{r^{(1)}}{1!}\,y^{(r-1)}(n-y)^{(1)}+\ldots+\frac{r^{(t)}}{t!}\,y^{(s)}(n-y)^{(t)}\right\} \quad \text{......(ii)}$$

$$=u(0)+u(1)+\ldots+u(t).$$

It is important to find a ready method of calculating the sum in (ii). This sum is what is meant by the probability integral for the hypergeometric. It is easily seen that, in limiting ourselves to the consideration of the sum of the first $(t+1)$ terms, instead of including also the sum of the last $(t+1)$ terms, we are really not losing generality, for the first sum would become the second if the values of p and q were interchanged, and the sum of all the terms is unity.

(2) The general plan and notation of this paper will follow fairly closely those of an earlier one, and we shall make use of certain of the results given there†. In particular; if A denote the tail of any frequency function $f(\xi)$ such that $f \geqq 0$ if $a \leqq \xi \leqq b$, and $f=0$ elsewhere, and if ‡

$$\Sigma_a^b f(\xi)\,\Delta\xi=1,\ A=\Sigma_a^\tau f(\xi)\,\Delta\xi=\{f(a)+f(a+\Delta\xi)+\ldots+f(\tau)\}\,\Delta\xi \quad \text{(iii)},$$

then we have, using the symbol "$\cong$" to mean "is approximately equal to," and Sheppard's § notation,

$$A \cong \phi(0)\left\{\frac{\Delta v}{2}+E\Delta v+\frac{\sigma}{2}\left(\frac{1-\alpha_x}{z_x}\right)\right\},$$

---

* Cf. W. P. Elderton, *Frequency Curves and Correlation*, Layton, London, 1906, p. 37.

† B. H. Camp, "Probability Integrals for the Point Binomial," *Biometrika*, Vol. XVI. pp. 163—171.

‡ The reader should note that, in the notation to be used, Σ_a^τ is a sum which includes the values at the limits a and τ.

§ *Tables for Statisticians and Biometricians*, 2nd Edition, Part I. 1924, pp. 2—11.

where, in the order suitable for computation,

$$v = \tau - \xi,\ \ \phi(v) = f(\tau - v),\ \ \Delta v = \Delta\xi,\ \ Q = \frac{\phi(\Delta v)}{\phi(0)},\ \ R = \sqrt{Q\,\frac{\phi(\Delta v)}{\phi(2\Delta v)}},$$

$$c_1 = -\log_e QR,\ \ c_2 = \log_e R,\ \ \frac{1}{\sigma} = \sqrt{2c_2}\,\Delta v,\ \ x = c_1\sigma,$$

$$E \cong \frac{c_1}{12}\left(1 + \frac{3 - x^2}{60\sigma^2} + \frac{15 - 10x^2 + x^4}{2520\sigma^4}\right) \ldots\ldots\ldots\ldots\ldots\ldots\text{(iv)}.$$

Sometimes certain terms of (iv) may be neglected, and then

$$A \cong \frac{\phi(0)\,\Delta v}{1 - Q^{\Delta v}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(v)}.$$

(3) *Special Case* (*Q small*). It is only when Q is small that (v) suffices, but when the stump is far from the mode this often occurs. We may then apply (v) directly to (ii) and obtain

$$S_{t+1}(y) \cong \frac{u(t)}{1 - Q(t)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(vi)},$$

where $u(t)$ is the first term in the stump, viz.,

$$u(t) = \frac{r^{(t)}}{n^{(r)}\,t!}\,y^{(s)}\,(n-y)^{(t)},\ \text{ and }\ Q(t) = \frac{t\,(y-s)}{(s+1)\,(n-y-t+1)}.$$

Sometimes it is worth while computing a few terms of the tail before applying (vi). When three extra terms are computed, $S_{t+1}(y)$ takes the form

$$S_{t+1}(y) = u(t-3)\left\{\frac{1}{1 - Q(t-3)} + \frac{1}{Q(t-2)} + \frac{1}{Q(t-2)}\cdot\frac{1}{Q(t-1)}\right.$$
$$\left. + \frac{1}{Q(t-2)}\cdot\frac{1}{Q(t-1)}\cdot\frac{1}{Q(t)}\right\} \ldots\ldots\text{(vi } bis\text{)};$$

where

$$Q(t-3) = \frac{(t-3)\,(y-s-3)}{(s+4)\,(n-y-t+4)},\qquad \frac{1}{Q(t-2)} = \frac{(s+3)\,(n-y-t+3)}{(t-2)\,(y-s-2)},\ \text{etc.}$$

(4) *General Case.* In order to be able to use for the general case methods analogous to those employed for the point binomial, it is necessary to show that an analogous transformation can be made of the sum S_{t+1}. The result will not be, as in the case of the point binomial, an integral, but instead another sum. However, this new sum will be found to possess a certain advantage due to the fact that the point binomial integral is its limiting form.

First, let us difference (ii) with respect to y, taking $\Delta y = 1$, and noting the formulae:

$$\Delta\,(n-y)^{(t)} = -t\,(n-y-1)^{(t-1)},$$
$$\Delta y^{(s)}\,(n-y)^{(t)} = s y^{(s-1)}\,(n-y)^{(t)} - t\,(y+1)^{(s)}\,(n-y-1)^{(t-1)};$$
$$\Delta S_{t+1}(y) = \frac{1}{n^{(r)}}\left\{r y^{(r-1)} + \frac{r^{(2)}}{1!}\,y^{(r-2)}\,(n-y)^{(1)} + \frac{r^{(3)}}{2!}\,y^{(r-3)}\,(n-y)^{(2)} + \ldots\right.$$
$$+ \frac{r^{(t+1)}}{t!}\,y^{(s-1)}\,(n-y)^{(t)} - \frac{r^{(1)}}{0!}\,(y+1)^{(r-1)}\,(n-y-1)^{(0)} - \frac{r^{(2)}}{1!}\,(y+1)^{(r-2)}\,(n-y-1)^{(1)} - \ldots$$
$$\left. - \frac{r^{(t)}}{(t-1)!}\,(y+1)^{(s)}\,(n-y-1)^{(t-1)}\right\} \ldots\ldots\text{(vii)}.$$

Now write (vii) successively for $y=\alpha$, $y=\alpha+1, \ldots, y=\alpha+\beta$, and add. The result is

$$\begin{aligned}\Delta S_{t+1}(\alpha)+\Delta S_{t+1}(\alpha+1)+\ldots+\Delta S_{t+1}(\alpha+\beta)&\\ =\frac{1}{n^{(r)}}\frac{r^{(t+1)}}{t!}\Sigma_{\xi=\alpha}^{\xi=\alpha+\beta}\xi^{(s-1)}(n-\xi)^{(t)}(\Delta\xi=1)&\end{aligned}$$

$$-\frac{1}{n^{(r)}}\left[\frac{r^{(t)}}{(t-1)!}\xi^{(s)}(n-\xi)^{(t-1)}+\frac{r^{(t-1)}}{(t-2)!}\xi^{(s+1)}(n-\xi)^{(t-2)}+\ldots+\frac{r^{(1)}}{0!}\xi^{(r-1)}(n-\xi)^{(0)}\right]_{\xi=\alpha}^{\xi=\alpha+\beta+1} \quad \ldots\ldots\ldots\text{(viii)}.$$

In this let us now put $\alpha=s-1$, $\alpha+\beta=y-1$. Then the left hand side of (viii) equals $S_{t+1}(y)$, for by (ii) $S_{t+1}(s-1)=0$. On the right the expression in brackets vanishes at the lower limit $\xi=s-1$, for each term involves $(s-1)$ with an index greater than $s-1$, omitting the trivial case where $t=0$. Thus we have

$$S_{t+1}(y)=\frac{1}{n^{(r)}}\frac{r^{(t+1)}}{t!}\Sigma_{\xi=s-1}^{\xi=y-1}\xi^{(s-1)}(n-\xi)^{(t)}\Delta\xi+\theta(y),$$

where

$$\theta(y)=-\frac{1}{n^{(r)}}\left\{\frac{r^{(t)}}{(t-1)!}y^{(s)}(n-y)^{(t-1)}+\frac{r^{(t-1)}}{(t-2)!}y^{(s+1)}(n-y)^{(t-2)}+\ldots+\frac{r^{(1)}}{0!}y^{(r-1)}(n-y)^{(0)}\right\} \quad \ldots\ldots\ldots\text{(ix)}.$$

In order to sum the series $\theta(y)$, we may repeat the method just used. Difference with respect to y, getting $\Delta\theta(y)$.

Then write $\Delta\theta(y)$ for $y=\alpha$, $y=\alpha+1, \ldots y=\alpha+\beta$, and add, obtaining

$$\Delta\theta(\alpha)+\Delta\theta(\alpha+1)+\ldots+\Delta\theta(\alpha+\beta).$$

If we then substitute $\alpha=s-1$, $\alpha+\beta=y-1$, we obtain $\theta(y)$, for $\theta(s-1)=0$, as already noted. The details, though somewhat lengthy, offer no difficulty, and are omitted. We obtain, finally,

$$\theta(y)=-\frac{1}{n^{(r)}}\frac{r^{(t+1)}}{(t-1)!}\Sigma_{\xi=s-1}^{\xi=y-1}\xi^{(s-1)}(n-\xi)^{(t-1)}(\Delta\xi=1)+\theta_1(y),$$

where

$$\theta_1(y)=\frac{1}{n^{(r)}}\Sigma_{i=1}^{i=t-1}\frac{r^{(t+1-i)}}{(t-1-i)!}y^{(s-1+i)}(n-y)^{(t-1-i)}(\Delta i=1) \quad \ldots\ldots\ldots\text{(x)}.$$

Now $\theta_1(y)$ is exactly like $\theta(y)$, except for sign and certain changes in the indices. The result of continually repeating this process may therefore be inferred, and the final result is

$$S_{t+1}(y)=\frac{r^{(t+1)}}{n^{(r)}}\Sigma_{\xi=s-1}^{\xi=y-1}\xi^{(s-1)}\left(\frac{(n-\xi)^{(t)}}{t!}-\frac{(n-\xi)^{(t-1)}}{(t-1)!}+\ldots\pm 1\right) \quad \ldots\ldots\text{(xi)}.$$

The series in (xi) is in the form

$${}_mC_t-{}_mC_{t-1}+\ldots\pm{}_mC_0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xii)},$$

where ${}_mC_t$ is the number of combinations of m things taken t at a time, etc. In order to sum it, we write Pascal's theorem for t, $t-1$, etc., and add, as follows:

$$\begin{aligned}{}_mC_t &= {}_{m-1}C_t+{}_{m-1}C_{t-1},\\ -{}_mC_{t-1} &= -{}_{m-1}C_{t-1}-{}_{m-1}C_{t-2},\\ {}_mC_{t-2} &= {}_{m-1}C_{t-2}+{}_{m-1}C_{t-3},\\ &\vdots\qquad\quad\vdots\qquad\quad\vdots\\ \text{Sum} &= {}_{m-1}C_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xiii)}.\end{aligned}$$

Insert this result in (xi):

$$S_{t+1}(y) = \frac{r^{(t+1)}}{n^{(r)}\,t!} \Sigma_{\xi=s-1}^{\xi=y-1} \xi^{(s-1)} (n-\xi-1)^{(t)} \quad (\Delta\xi = 1) \quad \text{.........(xiv)}.$$

This is the formula for the hypergeometric which is analogous to (xiv) in my paper on the Point Binomial. It is of the form (iii), and, if we now apply to it the approximation (iv), which, as in my previous paper, is expected to be good only when s and t are large, we have as

METHOD I (*Indirect Approximation*): $S_{t+1}(y) = u(t)\frac{s}{y}\rho$, where $u(t)$ is as in (vi), and, in the order suitable for computation,

$$Q(t) = \frac{(y-s)(n-y+1)}{(y-1)(n-y-t+1)}, \quad R = \sqrt{Q\frac{(y-2)(n-y-t+2)}{(y-s-1)(n-y+2)}},$$

$$c_1 = -\log_e QR, \quad c_2 = \log_e R, \quad \frac{1}{\sigma} = \sqrt{2c_2}, \quad x = c_1\sigma,$$

$$E \text{ as in (iv)}, \quad \rho = \tfrac{1}{2} + \frac{\sigma}{2}\left(\frac{1-\alpha_x}{z_x}\right) + E \quad \text{.....................(xv)}.$$

As in the case of the point binomial, it is necessary that $c_1 \geqq 0$, and this can always be secured by a proper statement of the problem initially. That is, if c_1 should ever come out negative, one would know that the tail of the hypergeometric as chosen had produced in (xiv) a function of ξ whose mode was within the limits of the summation indicated in that formula. The original problem should therefore be restated so that the opposite end of the hypergeometric is selected as the tail, and then c_1 will be found to be positive. The mode of the function of ξ that occurs is easily found to be at the point $\xi = n(s-1)/(r-1) - 1$.

Example 1. Find the sum of the first 46 terms of the hypergeometric for which $n = 360$, $r = 90$, $p = \frac{1}{3}$.

Here $\quad y = 120, \; t = 45, \; n - y = 240, \; Q = \frac{(75)(241)}{(196)(119)} = \cdot 774\ 9528,$

$$R = \sqrt{Q\frac{(118)(197)}{(74)(242)}}, \quad c_1 = 0{\cdot}251\ 9864, \quad c_2 = 0{\cdot}002\ 9662\ 102,$$

$$\frac{1}{\sigma} = 0{\cdot}077\ 0222, \quad x = 3{\cdot}271\ 607, \quad E = 0{\cdot}020\ 983, \quad \rho = 4{\cdot}1920,$$

$$S = 1{\cdot}5720, \quad u(t) = 0{\cdot}0001\ 1553;$$

and the true value of this sum, found by computing and adding the terms, is 0·0001 1525.

(5) METHOD II (*Direct Approximation*): As in the case of the point binomial one may approximate to the sum of the first $(t+1)$ terms of a hypergeometric by applying the formulae of Section 2 (iv) directly to the series. The formulae resulting from this process will now be given, and then, after a few illustrative examples, the relative merits of the two methods will be discussed:

$$S_{t+1} = u(t), \quad Q = \frac{t(y-s)}{(s+1)(n-y-t+1)}, \quad R = \sqrt{Q\frac{(s+2)(n-y-t+2)}{(t-1)(y-s-1)}},$$

$$c_1, c_2, \sigma, x, E, \text{ and } \rho \text{ as in (xv)} \quad \text{...........(xvi)}.$$

Example 2. Find the sum in Example 1 by (xvi). We obtain

$$c_1 = 0{\cdot}951\,358,\quad c_2 = 0{\cdot}031\,2456,\quad \frac{1}{\sigma} = 0{\cdot}249\,982,\quad x = 3{\cdot}80570,\quad E = 0{\cdot}078\,341,$$

$$\rho = 1{\cdot}56828,\quad S = 0{\cdot}000\,115\,253, \text{ and the true value is } 0{\cdot}000\,115\,254.$$

(6) The above examples have been worked with all the precision necessary to exhibit the closeness of approximation of which the formulae are capable. This has involved the use of inconvenient tables* to find $\frac{1}{2}(1-\alpha)/z$, and the necessity of finding the constants to several decimal places. For many purposes, however, as great accuracy as that obtained, for example, by (xvi), is not required. What is especially needed is a method which will give results in error by at most one or two per cent., and which will involve a minimum of computation. This last condition is very important, for we are here considering approximate formulae, and one always has the option of discarding the formulae and computing the several terms of the series outright. The approximate formulae, therefore, must not be too cumbersome to employ readily. Both of the methods given above admit of some modification, which will now be noted, and the succeeding examples will illustrate the accuracy which may be obtained with a minimum of effort.

First write, in both cases, $E \cong c_1/12$, and so:

Indirect Method I: $$S_{t+1} = \frac{s}{y}\,u(t)\left\{\frac{1}{2} + \frac{c_1}{12} + \frac{\sigma}{2}\,\frac{(1-\alpha_x)}{z_x}\right\} \quad\ldots\ldots\ldots\ldots\ldots\ldots(\text{xvii}),$$

Direct Method II: $$S_{t+1} = u(t)\left\{\frac{1}{2} + \frac{c_1}{12} + \frac{\sigma}{2}\,\frac{(1-\alpha_x)}{z_x}\right\} \ldots\ldots\ldots\ldots\ldots\ldots(\text{xviii}).$$

In both cases, find Q and R^2 by a computing machine, and then $\log_e Q$ and $\log_e R^2$ from a five place table of natural† logarithms. Obtain σ to four significant figures by the use of tables of reciprocals and square roots, and find the ratio $\frac{1}{2}(1-\alpha)/z$ from Sheppard's table‡.

Example 3. Do Example 1 as just indicated. By Method I, $S = 0{\cdot}000\,116$; by Method II, $S = 0{\cdot}000\,115$; the true value is $0{\cdot}000\,115$.

Example 4. Let $n = 900$, $r = 450$, $p = \frac{2}{3}$, $t = 120$. By Method I, $S = 0{\cdot}000\,0145$; by Method II, $S = 0{\cdot}000\,0144$; the true value is $0{\cdot}000\,0145$.

Example 5. Let $n = 900$, $r = 450$, $p = \frac{2}{3}$, $t = 110$. By Method I, $S = 1{\cdot}00 \times 10^{-8}$; by Method II, $S = 1{\cdot}01 \times 10^{-8}$; the true value is $1{\cdot}02 \times 10^{-8}$.

Example 6. Let $n = 240$, $r = 90$, $p = \frac{1}{3}$, $t = 45$. By Method I, $S = 0{\cdot}0000\,2285$; by Method II, $S = 0{\cdot}0000\,2282$; the true value is $0{\cdot}0000\,2282$.

Example 7. Let $n = 360$, $r = 90$, $p = \frac{1}{3}$, $t = 72$. By Method I, $S = 0{\cdot}002\,783$; by Method II, $S = 0{\cdot}002\,756$; the true value is $0{\cdot}002\,754$.

---

* Burgess, J., *Transactions of the Royal Society of Edinburgh*, 1897, Vol. XXXIX. pp. 257—321.

† E.g., B. O. Peirce, *A Short Table of Integrals*, 2nd Ed., 1910, Ginn and Co., Boston, p. 130.

‡ Unfortunately, this table is so arranged that the ratio is given poorly in the range, $4 \leqq x < 4{\cdot}50$. A tabulation of the ratio itself would be very helpful here, as with the point binomial approximation. Such a table, accurate to five significant figures, is now in course of preparation.

Example 8. Let $n=900$, $r=90$, $p=\frac{1}{9}$, $t=72$. By Method I, $S=0{\cdot}006\ 535$; by Method II, $S=0{\cdot}006\ 428$; the true value is 0·006 428.

Example 9. Let $n=900$, $r=90$, $p=\frac{1}{3}$, $t=45$. By Method I, $S=0{\cdot}000\ 422$; by Method II, $S=0{\cdot}000\ 414$; the true value is 0·000 414.

A critical survey of these examples will indicate that the direct method (II) gives better results. There is no essential difference between the two methods as regards difficulty of computation. Is there, therefore, justification for the development of the indirect method? This question may be answered after a brief consideration of the point binomial case. There the indirect method was preferred, not because it gave closer results—indeed, in the published examples it gave less close results—but because it gave results which were close enough, and it involved less computation. It was, for large values of r (n in that paper), the more *efficient* approximation, efficiency being understood to depend partly on the closeness of the approximation, and partly on the simplicity of the formulae. Now, if we let n become infinite in our present indirect formulae, they will approach as limits the corresponding indirect formulae for the point binomial; the sum (xiv) of this paper will approach the integral (xiv) of that one. Therefore, in border line cases, that is, where n/r is large, it is to be expected, and it is in fact true that, without serious loss of accuracy, the formulae of this method may be simplified, so as to render them analogous to the point binomial formulae, and so as to make them easier to use in computing. On the other hand, although the direct method formulae are susceptible of an analogous transformation, going over into the direct method formulae for the point binomial, they do not result in a corresponding simplification. The simplified formulae will now be given and illustrated.

(7) *Use of Indirect Method when n/r is large.* As before, use tables of reciprocals and squares and find the following constants accurately to four significant figures:

$$c_1=\frac{s-1}{y-1}-\frac{t}{n-y-t+1}+\tfrac{1}{2}\left[\left(\frac{s-1}{y-1}\right)^2+\left(\frac{t}{n-y-t+1}\right)^2\right],$$

$$\frac{1}{\sigma^2}=\frac{-1}{y-1}+\frac{1}{n-y-t+1}-\frac{1}{n-y+2}+\frac{1}{y-s-1}, \quad x=c_1\sigma.$$

Then $$S_{t+1}=u(t)\frac{s}{y}\left(\tfrac{1}{2}+\frac{\sigma}{2}\frac{1-\alpha_x}{z_x}\right) \qquad \text{(xix)}.$$

Example 10. Use the above method in Example 9. We obtain $n/r=10$, $S=0{\cdot}000\ 424$; by I, Ex. 9, $S=0{\cdot}000\ 422$: the true value is 0·000 414. Thus there is no essential difference in the approximation given by Method I when these simpler formulae for the constants are used. Since r is only 90, neither result is very close to the true value. As with the point binomial, the success of the method depends on r being large*. This method is especially valuable when r is much larger, as in the following case.

---

* I have intentionally included in both my papers examples where the methods do not apply very well, in order that the reader may have some idea of the range of their usefulness.

Example 11. Let $n = 4500$, $r = 450$, $p = \frac{2}{3}$, $t = 120$. Here also $n/r = 10$. $S = 0{\cdot}000\,7965$; the true value is 0·000 7980. The proportional error is only about 0·19 of one per cent., one-twelfth as large as that in Example 10. Also, the last place is in some doubt, due to the inadequacy of the tables for the ratio, $\frac{1}{2}(1 - \alpha)/z$.

(8) *Summary.* In order that the reader may have the benefit of such experience as the author has had in computing from the above formulae, some suggestions as to when the various formulae should be used will now be given. These suggestions are admittedly both tentative and proximate, for the experience on which they are based is not exhaustive, and because, when the accuracy of a formula depends on four variables (n, r, p, s), it is seldom possible to frame a simple, rigorous rule of procedure.

(*a*) If greater accuracy is desired than can be obtained by the abbreviated methods of Section (6), it is probably not worth while using the formulae at all unless r is at least as large as 300. This would not be true if better auxiliary tables were available*, but in our present state it is probably better to compute the various terms of the tail by successive multiplications, beginning with the term at the stump.

(*b*) If $pq > \frac{1}{10}$, and $r < 400$, use Method II of Section (6). If $pq > \frac{1}{10}$ and $r > 400$, use Method II of Section (6) for values of $n < 10r$, Method I as modified in Section (7) for higher values of n. The proportional errors will probably not exceed one- or two-tenths of one per cent. in the extreme cases, and will usually be much smaller. They grow small as pq and r are increased.

(*c*) Formula (vi *bis*), or even (vi), is always to be used, however, when the Q of Method II, which is the same as the Q of (vi), is small, say less than one-third. A case in which Q is a little greater than one-third (0·3743) is given in Examples 2 and 3. The error in S as given by the method of Example 2 is 0·000 000 001; by II of Example 3 it is 0·000 000 02; by (vi *bis*) it is 0·000 000 19; and by (vi) it is 0·000 002 26. The proportional error by (vi *bis*) is 0·0017, and by (vi) is 0·02.

Finally, it is a pleasure to record my indebtedness to Professor Karl Pearson for proposing the problem of this paper and for helpful suggestions made during the course of its preparation.

---

* Cf. footnote ‡ on p. 65 above.

DETERMINATION OF $\int_0^\theta \cos^{n+1}\theta\, d\theta$ FOR LARGE VALUES OF n, AND ITS APPLICATION TO THE PROBABILITY INTEGRAL OF SYMMETRICAL FREQUENCY CURVES.

BY JOHN WISHART, M.A., B.Sc.

THE Complete Integral $\int_0^{\pi/2} \cos^{n+1}\theta\, d\theta$ has been tabled from $n = -1$ to 103*. The Incomplete Integral

$$\int_0^\theta \cos^{n+1}\theta\, d\theta = \tfrac{1}{2}\int_0^x (1-x)^{\frac{1}{2}n} x^{-\frac{1}{2}}\, dx \quad \text{on putting} \quad x = \sin^2\theta$$

$$= \tfrac{1}{2} B_x\left(\tfrac{1}{2}n + 1, \tfrac{1}{2}\right), \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

and so can be found from Tables of the Incomplete B-function, which are in progress. But such tables must of necessity be limited, stopping, say, at $p = q = 50$. It still remains a problem then to determine the integral under consideration for large values of n. It is the object of this paper to show how our function can be evaluated to a very great degree of accuracy by expanding in Incomplete Normal Moment functions. The method was first suggested by Professor Pearson as a means of evaluating the incomplete Γ- and B-functions†, but it was found after tables of the moment functions had been made that it was not accurate enough when n was small. Illustrations of the method in the case of the Incomplete B-function appear in *Tracts for Computers*, No. VII, but the accuracy there achieved leaves something to be desired.

As will be shown later, the curve of which (1) is the integral is one of the well-known Pearson frequency types, and hence what is generally required is the ratio of the incomplete to the complete integral.

Let
$$\tfrac{1}{2} + I_\theta(n+1) = \frac{\int_{-\pi/2}^{\theta} \cos^{n+1}\theta\, d\theta}{\int_{-\pi/2}^{+\pi/2} \cos^{n+1}\theta\, d\theta}.$$

Then
$$I_\theta(n+1) = \frac{\int_0^\theta \cos^{n+1}\theta\, d\theta}{2\int_0^{\pi/2} \cos^{n+1}\theta\, d\theta} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

$$\frac{\int_0^y (1-y^2)^{\frac{1}{2}n}\, dy}{2\int_0^1 (1-y^2)^{\frac{1}{2}n}\, dy} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

on putting $y = \sin\theta$.

---

* *Biometrika*, Vol. XI. p. 377

† *Biometrika*, Vol. VI. pp. 66—8.

Now $$\int_0^1 y^{2m-1}(1-y^2)^{p-1}\,dy = \tfrac{1}{2}\cdot\frac{\Gamma(m)\,\Gamma(p)}{\Gamma(m+p)}.$$

In this put $m=\tfrac{1}{2}$ and $p=\tfrac{1}{2}n+1$.

$$\therefore\ I_\theta(n+1)=\frac{1}{\sqrt{\pi}}\cdot\frac{n+1}{n}\cdot\frac{\Gamma\{\frac{1}{2}(n+1)\}}{\Gamma(\frac{1}{2}n)}\int_0^y (1-y^2)^{\frac{1}{2}n}\,dy.$$

Again $$\Gamma(\tfrac{1}{2}n)\,\Gamma\{\tfrac{1}{2}(n+1)\}=\frac{\sqrt{\pi}}{2^{n-1}}\,\Gamma(n)$$

and we shall expand the Γ-functions by Stirling's theorem. Thus:

$$I_\theta(n+1)=\frac{1}{2^{n-1}}\left(1+\frac{1}{n}\right)\frac{\Gamma(n)}{\{\Gamma(\frac{1}{2}n)\}^2}\int_0^y e^{-\frac{1}{2}n\left(y^2+\frac{1}{2}y^4+\frac{1}{3}y^6+\frac{1}{4}y^8+\ldots\right)}\,dy \ \ldots\ldots\ldots(4)$$

$$=\left(1+\frac{1}{n}\right)\left(1-\frac{1}{4n}+\frac{1}{32n^2}+\frac{5}{128n^3}\right)\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{x^2}{2}}\,.\,e^{-\frac{x^4}{4n}-\frac{x^6}{6n^2}-\frac{x^8}{8n^3}}\,dx,$$

on putting $x=\sqrt{n}\,.\,y$ in the integral. Terms in $\frac{1}{n^4}$ and higher powers are neglected.

$$\therefore\ I_\theta(n+1)=\left(1+\frac{3}{4n}-\frac{7}{32n^2}+\frac{9}{128n^3}\right)\frac{1}{\sqrt{2\pi}}$$
$$\times\int_0^x e^{-\frac{x^2}{2}}\left[1-\frac{x^4}{4n}-\frac{x^6}{6n^2}+x^8\left(\frac{1}{32n^2}-\frac{1}{8n^3}\right)+\frac{x^{10}}{24n^3}-\frac{x^{12}}{384n^3}\right]dx.$$

Now the rth Incomplete Normal Moment function is

$$m_r(x)=\frac{1}{\sqrt{2\pi}}\int_0^x x^r e^{-\frac{1}{2}x^2}\,dx\Big/(r-1)(r-3)\ldots\ldots 2 \text{ or } 1,$$

and so in terms of these

$$I_\theta(n+1)=\left(1+\frac{3}{4n}-\frac{7}{32n^2}+\frac{9}{128n^3}\right)m_0(x)-\left(\frac{3}{4n}+\frac{9}{16n^2}-\frac{21}{128n^3}\right)m_4(x)$$
$$-\left(\frac{5}{2n^2}+\frac{15}{8n^3}\right)m_6(x)+\left(\frac{105}{32n^2}-\frac{1365}{128n^3}\right)m_8(x)+\frac{315}{8n^3}m_{10}(x)-\frac{3465}{128n^3}m_{12}(x)\ \ldots(5)$$
$$=c_0\,.\,m_0(x)-c_4\,.\,m_4(x)-c_6\,.\,m_6(x)+c_8\,.\,m_8(x)+c_{10}\,.\,m_{10}(x)-c_{12}\,.\,m_{12}(x)\ \ldots\ldots(6),$$

where

$$\left.\begin{aligned}
c_0&=1+\frac{\cdot 75}{n}-\frac{\cdot 21875}{n^2}+\frac{\cdot 0703125}{n}\\
c_4&=\frac{\cdot 75}{n}+\frac{\cdot 5625}{n^2}-\frac{\cdot 1640625}{n^3}\\
c_6&=\frac{2\cdot 5}{n^2}+\frac{1\cdot 875}{n^3}\\
c_8&=\frac{3\cdot 28125}{n^2}-\frac{10\cdot 6640625}{n^3}\\
c_{10}&=\frac{39\cdot 375}{n^3}\\
c_{12}&=\frac{27\cdot 0703125}{n^3}
\end{aligned}\right\}\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(7).$$

In this expression $x = \sqrt{n}\sin\theta$. $m_0(x)$ is got from the table of $\frac{1}{2}(1+\alpha)$* by subtracting ·5, and the functions $m_r(x)$ from $r = 1$ to 10 are tabled to seven figures*. A supplementary table giving $m_{12}(x)$ appears below. It was calculated by integrating twice by parts and combining the results, which gives the relation

$$m_{12}(x) = \tfrac{1}{2}\{m_8(x) + m_{10}(x)\} - \frac{x^9}{1890}\left(1 + \frac{2x^2}{11}\right)z_x,$$

where z_x is the ordinate of the normal curve, given by Sheppard's tables. In a few cases where the last figure was in doubt, the integration was carried further, and the function expressed in terms of $m_0(x)$, known to more places than seven. Second differences only are necessary for interpolation into this table, for although the third differences are considerable, none is greater than ·0006, and this, when combined with the appropriate coefficient, is insignificant to the order of accuracy we desire.

This result (equation (6)) enables the function $I_\theta(n+1)$ to be calculated with great rapidity when once the coefficients c_0 etc. are known, and provided that n is reasonably large. Table I gives the values of these coefficients for values of n from 100 to 400, proceeding by twos, i.e. for values which would come outside the Tables of the Incomplete B-function. For still higher values of n, c_0, c_4, c_6 and c_8 can be calculated from (7), dropping the terms in $\frac{1}{n^3}$. It is found that for n above 400, the terms in $\frac{1}{n^3}$ are insignificant in the seventh place, and to a first approximation

$$I_\theta(n+1) = \left(1 + \frac{3}{4n} - \frac{7}{32n^2}\right)m_0(x) - \left(\frac{3}{4n} + \frac{9}{16n^2}\right)m_4(x) - \frac{5}{2n^2}m_6(x) + \frac{105}{32n^2}m_8(x) \ldots (8).$$

Now the equation to the symmetrical frequency curve of limited range (Pearson, Type IIa) is

$$z = z_0\left(1 - \frac{x^2}{a^2}\right)^m \dagger.$$

If in this we put $y = \frac{x}{a}$ and $m = \frac{1}{2}n$ it becomes

$$z = z_0(1 - y^2)^{\frac{1}{2}n},$$

and hence the probability integral of this type of curve is

$$\frac{\int_{-1}(1-y^2)^{\frac{1}{2}n}\,dy}{\int_{-1}^{+1}(1-y^2)^{\frac{1}{2}n}\,dy} = \tfrac{1}{2} + I_\theta(n+1),$$

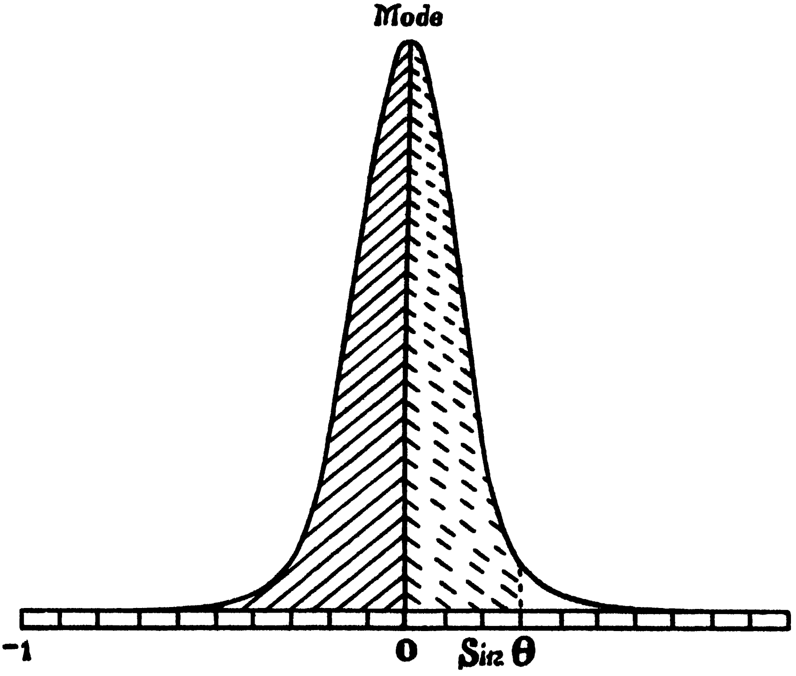
from (3).

So that the function determined by equation (6) is the probability integral of the Type IIa curve, when added to $\frac{1}{2}$. The diagram illustrates the form the curve takes for large values of n, and the probability integral is the ratio of the shaded portion of the area to the whole area. The examples which will now be given in illustration of the theory will show that the approximation (6) determines $I_\theta(n+1)$ to seven figure accuracy up to a distance from the mode (taken as origin) of three times the standard deviation. This is very nearly the whole area. The value of the "tail" is given by $\frac{1}{2} - I_\theta(n+1)$.

---

* *Tables for Statisticians*, pp. 2—8 and 22—3.

† The type $z = z_0\left(1 + \frac{x^2}{a^2}\right)^{-m}$ reduces to the same integral form if we take $x/a = \tan\theta$; no further probability integral is required.



Example I. To evaluate $I_\theta(101)$.

$$\sigma = \frac{1}{\sqrt{n+3}} = \cdot 0985.$$

(a) Let us integrate from the mode to $\theta = \sin^{-1} \cdot 3$.

Then $x = \sqrt{n} \sin\theta = 3$. We have at once from equation (6) and using Table I

$$\begin{array}{rlrl} I_{\cdot 3}(101) = & 1\cdot 0074\ 782\ m_0(3) & - & \cdot 0075\ 561\ m_4(3) \\ & -\cdot 0002\ 519\ m_6(3) & + & \cdot 0003\ 175\ m_8(3) \\ & +\cdot 0000\ 394\ m_{10}(3) & - & \cdot 0000\ 271\ m_{12}(3) \\ = & \cdot 5023\ 7911 & - & \cdot 0033\ 6600 \\ & 8933 & & 9413 \\ & 745 & & 308 \\ \hline & \cdot 5024\ 7589 & & \cdot 0034\ 6321 \\ & 34\ 6321 & & \\ \hline & \cdot 4990\ 1268 & & \end{array}$$

$$\therefore\ I_{\cdot 3}(101) = \cdot 4990\ 127 \text{ to seven places.}$$

As a check on this result, a value, correct to at least seven figures, was obtained from 16 terms of the "parts" reduction formula. Since

$$\int_0^y (1-y^2)^{\frac{1}{2}n}\, dy = \int_0^y (1+y)^{\frac{1}{2}n}(1-y)^{\frac{1}{2}n}\, dy$$

the function we are evaluating is seen to be equal to $\dfrac{B_y(51, 51)}{B(51, 51)}$, taken from the mode as origin, and when we remember that our range is ± 1 from the mode and the integral of the "parts" reduction formula is measured from the start of the curve, we see that

$$I_{\cdot 3}(101) = \frac{1}{2} - \frac{B_{\cdot 35}(51, 51)}{B(51, 51)}$$

$$= \frac{1}{2} - \frac{\Gamma(102)}{\Gamma(52)\,\Gamma(51)}(\cdot 35)^{51}(\cdot 65)^{50}\left(1 + \frac{50}{52}\cdot\frac{7}{13}\left\{1 + \frac{49}{53}\cdot\frac{7}{13}\left\{\cdots\cdots\left\{1 + \frac{36}{66}\cdot\frac{7}{13}\right\}\right\}\cdots\right\}\right),$$

the Remainder Integral being insignificant to this order of accuracy,

$$= \frac{1}{2} - \frac{\Gamma(102)}{\Gamma(52)\,\Gamma(51)}(\cdot 35)^{51}(\cdot 65)^{50} \times 1\cdot 9984\ 3246.$$

This expression was evaluated using Tables of the Logarithms of the Gamma-Function* and Briggs' Logarithms to 12 places.

We have

$$\begin{array}{r}
159\cdot 974\ 325\ 0285 \\
\overline{24}\cdot 747\ 470\ 2618 \\
1\overline{0}\cdot 645\ 667\ 8322 \\
0\cdot 300\ 689\ 4751 \\
\hline
127\cdot 668\ 152\ 5976 \\
130\cdot 673\ 719\ 9211 \\
\hline
\overline{4}\cdot 994\ 432\ 6765
\end{array}$$

$\therefore\ I_{\cdot 3}(101) = \frac{1}{2} - \cdot 0009\ 8726 = \cdot 4990\ 127$ to seven places,

which shows that our value by moment functions is correct to seven figures. A further attempt was made towards evaluating the integral, using Weddle's Rule (24 ordinates). This gave a value ·4990 128.

The actual value of $\int_0^\theta \cos^{101}\theta\, d\theta$, where $\theta = \sin^{-1}\cdot 3$, is, of course, obtained by multiplying the value found for $I_\theta(101)$ by $2\int_0^{\pi/2}\cos^{101}\theta\, d\theta$, which is tabled.

(*b*) Let us now integrate over a smaller range, say to $\theta = \sin^{-1}\cdot 2$. Then $x = \sqrt{n} \times \cdot 2 = 2$. The coefficients are the same as before, but we now require to look out the moment coefficients of $x = 2$. We find numerically:

$$\begin{array}{rr}
+\ \cdot 4808\ 1884 & -\ \cdot 0017\ 0233 \\
1406 & 2774 \\
59 & 12 \\
\hline
\cdot 4808\ 3349 & \cdot 0017\ 3019 \\
17\ 3019 & \\
\cline{1-1}
\cdot 4791\ 0330 &
\end{array}$$

$\therefore\ I_{\cdot 2}(101) = \cdot 4791\ 033$ to seven places, which is again found to be identical to this number of places with the result deduced by parts.

---

* *Tracts for Computers*, No. VIII.

Example II. Consider now the case of a higher power.

Let it be proposed to find $I_\theta(401)$. Now the point of inflection of the curve is given by $\sin\theta = y = \frac{1}{\sqrt{n-1}}$, so for the case considered in Example I the points of inflection are $\pm{\cdot}1$ from the mode, and in this example $\pm{\cdot}05$. Sometimes an approximate formula is found to be unreliable in the neighbourhood of a point of inflection. To examine this difficulty, let us take

(*a*) Integration to $\theta = \sin^{-1}{\cdot}05$, i.e. to a distance from the mode a little greater than the standard deviation, $\sigma = {\cdot}0498$.

Then
$$x = \sqrt{n}\sin\theta = 1.$$

Equation (8) may be used, the extra terms of (5) being found to be of no significance in the seventh place. We then have

$$\begin{aligned} I_{\cdot 05}(401) &= 1{\cdot}0018\ 736\ m_0(1) - {\cdot}0018\ 785\ m_4(1) \\ &\quad - {\cdot}0000\ 156\ m_6(1) + {\cdot}0000\ 205\ m_8(1) \\ &= \begin{array}{r} {\cdot}3419\ 8430 \\ 3520 \\ \hline {\cdot}3419\ 491 \end{array} \quad \begin{array}{r} -{\cdot}0000\ 3516 \\ 4 \\ \\ \end{array} \end{aligned}$$

$\cdot 3419\ 491$ to seven places,

the result to this number of places being, in fact, got from the first two terms only.

Two checks were here employed. Weddle's Rule, applied to 24 ordinates, gave the result ·3419 491. Then the integral was expanded and integrated term by term, as follows:

From (4)
$$\begin{aligned} I_\theta(n+1) &= \frac{1}{2^{n-1}}\left(1+\frac{1}{n}\right)\frac{\Gamma(n)}{\{\Gamma(\frac{1}{2}n)\}^2}\int_0^y (1-y^2)^{\frac{1}{2}n}\,dy \\ &= \frac{2{\cdot}005}{2^{400}}\,\frac{\Gamma(400)}{\{\Gamma(200)\}^2}\left[({\cdot}05) - \frac{200}{3}({\cdot}05)^3 + \frac{200\,.\,199}{2!\,5}({\cdot}05)^5 - \dots\right]. \end{aligned}$$

We need seven terms of the series for seven figure accuracy.

$$\therefore\ I_{\cdot 05}(401) = \frac{\Gamma(400)}{2^{400}\{\Gamma(200)\}^2} \times {\cdot}0857\ 675\ 241.$$

We have
$$\begin{array}{r} 866{\cdot}204\ 354\ 1864 \\ \bar{2}{\cdot}933\ 322\ 8727 \\ \hline 865{\cdot}137\ 677\ 0591 \\ 865{\cdot}603\ 715\ 5544 \\ \hline \bar{1}{\cdot}533\ 961\ 5047 \end{array} \qquad \begin{array}{r} 745{\cdot}191\ 717\ 2888 \\ 120{\cdot}411\ 998\ 2656 \\ \hline \\ \\ \\ \end{array}$$

$$\therefore\ I_{\cdot 05}(401) = {\cdot}3419\ 491 \text{ to seven places,}$$

so that all three methods yield, to this number of places, identical results.

(*b*) To integrate now over the range 3σ.

Let $\theta = \sin^{-1} \cdot 144$. $\therefore$ $x = 20 \times \cdot 144 = 2\cdot 88$.

The values of the moment functions of 2·88 were obtained by interpolation, using Everett's Central Difference Formula

$$z_\theta = \phi z_0 + \theta z_1 - \tfrac{1}{6}\theta\phi\,\{(\phi+1)\,\delta^2 z_0 + (\theta+1)\,\delta^2 z_1\}, \qquad \theta = \cdot 8,\ \phi = \cdot 2.$$

Consideration of the fourth differences throughout the table of moment functions shows us that, though in some cases they are considerable, the next term in Everett's Formula, taken in conjunction with the maximum coefficient of Table I, is negligible to our order of accuracy.

We then have

$$\begin{aligned} I_{\cdot 144}(401) &= 1\cdot 0018\ 736 \times \cdot 4980\ 11624 \\ &\quad - \cdot 0018\ 785 \times \cdot 4296\ 294 \\ &\quad - \cdot 0000\ 156 \times \cdot 3463\ 230 \\ &\quad + \cdot 0000\ 205 \times \cdot 2476\ 137 \\ &= \cdot 4981\ 373 \text{ to seven places.} \end{aligned}$$

The value found by Weddle's Rule was also ·4981 373, while integration term by term, a much more lengthy process this time, since 20 terms are required for seven figure accuracy, gave

$$\begin{aligned} I_{\cdot 144}(401) &= \frac{\Gamma(400)}{2^{400}\{\Gamma(200)\}^2} \times \cdot 1249\ 425\ 775 \\ &= \cdot 4981\ 373. \end{aligned}$$

Conclusion. The foregoing examples prove fairly conclusively that results, correct to seven places, can be obtained by expanding the function (2) in terms of the tabled seven figure moment functions. The first example would normally be obtained from a Table of the Incomplete *B*-function, when available, but it was selected as a case of extremely low indices. It is clear that the larger *n* is, the easier does the computation become, for then equation (5) reduces to the comparatively simple form of (8).

TABLE I.

Values of the Coefficients.

n	c_0	c_4	c_6	c_8	c_{10}	c_{12}
100	1·0074782	·0075561	·0002519	·0003175 −	·0000394	·0000271
102	1·0073320	74069	2421	3053	371	255 +
4	1·0071[illegible]4	72634	2328	2939	350 +	241
6	1·0070561	71254	2241	2831	331	227
8	1·0069257	69925 +	2158	2728	313	215 −
110	1·0068002	68645 +	2080	2632	296	203
112	1·0066790	67412	2006	2540	280	193
4	1·0065622	66221	1936	2453	266	183
6	1·0064493	65072	1870	2370	252	173
8	1·0063403	63962	1807	2292	240	165 −
120	1·0062348	62890	1747	2217	228	157
122	1·0061329	61852	1690	2146	217	149
4	1·0060342	60849	1636	2078	207	142
6	1·0059386	59877	1584	2013	197	135 +
8	1·0058461	58936	1535 −	1952	188	129
130	1·0057563	58024	1488	1893	179	123
132	1·0056693	57140	1443	1837	171	118
4	1·0055849	56283	1400	1783	164	113
6	1·0055029	55451	1359	1732	157	108
8	1·0054233	54643	1320	1682	150 −	103
140	1·0053460	53858	1282	1635 +	143	99
142	1·0052709	53095 +	1246	1590	138	95 −
4	1·0051978	52354	1212	1547	132	91
6	1·0051267	51633	1179	1505 +	127	87
8	1·0050576	50932	1147	1465 +	121	84
150	1·0049903	50250 −	1117	1427	117	80
152	1·0049248	49585 +	1087	1390	112	77
4	1·0048609	48938	1059	1354	108	74
6	1·0047987	48308	1032	1320	104	71
8	1·0047381	47693	1006	1287	100	69
160	1·0046790	47094	981	1256	96	66
162	1·0046213	46510	957	1225 +	93	64
4	1·0045651	45940	934	1196	89	61
6	1·0045101	45384	911	1167	86	59
8	1·0044566	44842	890	1140	83	57
170	1·0044042	44312	869	1114	80	55 +
172	1·0043531	43794	849	1088	77	53
4	1·0043031	43289	829	1064	75 −	51
6	1·0042543	42795 −	811	1040	72	50 −
8	1·0042066	42312	792	1017	70	48
180	1·0041599	41840	775 −	994	68	46
182	1·0041143	41378	758	973	65 +	45 −
4	1·0040696	40927	741	952	63	43
6	1·0040259	40485 −	726	932	61	42
8	1·0039832	40053	710	912	59	41
190	1·0039413	39629	695 +	893	57	39
192	1·0039003	39215 −	681	875 +	56	38
4	1·0038602	38809	667	857	54	37
6	1·0038208	38412	653	840	52	36
8	1·0037823	38022	640	823	51	35 −
200	1·0037445 +	37640	627	807	49	34

TABLE I (*continued*).

Values of the Coefficients.

n	c_0	c_4	c_6	c_8	c_{10}	c_{12}
202	1·0037075+	·0037266	·0000615−	·0000791	·0000048	·0000033
4	1·0036712	36900	603	776	46	32
6	1·0036356	36540	591	761	45+	31
8	1·0036007	36188	580	747	44	30
210	1·0035665−	35842	569	733	43	29
212	1·0035329	35502	558	719	41	28
4	1·0034999	35169	548	706	40	28
6	1·0034675+	34843	538	693	39	27
8	1·0034358	34522	528	680	38	26
220	1·0034046	34207	518	668	37	25+
222	1·0033739	33898	509	656	36	25−
4	1·0033439	33594	500−	644	35+	24
6	1·0033143	33296	491	633	34	23
8	1·0032853	33003	482	622	33	23
230	1·0032567	32715−	474	612	32	22
232	1·0032287	32432	466	601	32	22
4	1·0032011	32154	458	591	31	21
6	1·0031740	31881	450+	581	30	21
8	1·0031474	31612	443	571	29	20
240	1·0031212	31348	435+	562	28	20
242	1·0030954	31088	428	553	28	19
4	1·0030701	30832	421	544	27	19
6	1·0030452	30581	414	535+	26	18
8	1·0030206	30333	408	527	26	18
250	1·0029965+	30090	401	518	25+	17
252	1·0029727	29850+	395−	510	25−	17
4	1·0029494	29615−	389	502	24	17
6	1·0029264	29383	383	494	23	16
8	1·0029037	29154	377	487	23	16
260	1·0028814	28929	371	479	22	15+
262	1·0028594	28708	365+	472	22	15+
4	1·0028378	28490	360	465−	21	15−
6	1·0028165−	28275−	354	458	21	14
8	1·0027955−	28063	349	451	20	14
270	1·0027748	27855−	344	445−	20	14
272	1·0027544	27649	339	438	20	13
4	1·0027343	27447	334	432	19	13
6	1·0027145+	27248	329	426	19	13
8	1·0026950+	27051	324	420	18	13
280	1·0026758	26857	320	414	18	12
282	1·0026568	26666	315+	408	18	12
4	1·0026381	26478	311	402	17	12
6	1·0026197	26292	306	397	17	12
8	1·0026015+	26109	302	391	16	11
290	1·0025836	25929	298	386	16	11
292	1·0025659	25751	294	381	16	11
4	1·0025485−	25575+	290	375+	15+	11
6	1·0025313	25402	286	370	15+	10
8	1·0025143	25231	282	365+	15−	10
300	1·0024976	25062	278	361	15−	10

TABLE I (*continued*).

Values of the Coefficients.

n	c_0	c_4	c_6	c_8	c_{10}	c_{12}
302	1·0024810	·0024896	·0000275 –	·0000356	·0000014	·0000010
4	1·0024647	24732	271	351	14	10
6	1·0024483	24570	268	347	14	9
8	1·0024328	24410	264	342	13	9
310	1·0024171	24252	261	338	13	9
312	1·0024016	24096	257	334	13	9
4	1·0023863	23942	254	329	13	9
6	1·0023712	23790	251	325	12	9
8	1·0023563	23640	248	321	12	8
320	1·0023416	23492	245 –	317	12	8
322	1·0023271	23346	242	313	12	8
4	1·0023127	23202	239	309	12	8
6	1·0022986	23059	236	306	11	8
8	1·0022846	22918	233	302	11	8
330	1·0022707	22779	230	298	11	8
332	1·0022571	22641	227	295 –	11	7
4	1·0022435 +	22505 +	225 –	291	11	7
6	1·0022302	22371	222	288	10	7
8	1·0022170	22239	219	284	10	7
340	1·0022040	22107	217	281	10	7
342	1·0021911	21978	214	278	10	7
4	1·0021784	21850 –	212	275 –	10	7
6	1·0021658	21723	209	272	10	7
8	1·0021534	21598	207	268	9	6
350	1·0021411	21474	205 –	265 +	9	6
352	1·0021289	21352	202	262	9	6
4	1·0021169	21231	200	259	9	6
6	1·0021050 +	21112	198	257	9	6
8	1·0020933	20994	195 +	254	9	6
360	1·0020816	20877	193	251	8	6
362	1·0020702	20761	191	248	8	6
4	1·0020588	20647	189	245 +	8	6
6	1·0020475 +	20534	187	243	8	6
8	1·0020364	20422	185 –	240	8	5 +
370	1·0020254	20311	183	238	8	5 +
372	1·0020145 +	20202	181	235 +	8	5 +
4	1·0020038	20094	179	233	8	5 +
6	1·0019931	19987	177	230	7	5 +
8	1·0019826	19881	175 +	228	7	5 +
380	1·0019722	19776	173	225 +	7	5 –
382	1·0019619	19672	172	223	7	5 –
4	1·0019516	19569	170	221	7	5 –
6	1·0019415 +	19468	168	218	7	5 –
8	1·0019315 +	19367	166	216	7	5 –
390	1·0019216	19268	165 –	214	7	5 –
392	1·0019118	19169	163	212	7	4
4	1·0019021	19072	161	210	6	4
6	1·0018925 +	18975 +	160	208	6	4
8	1·0018830	18880	158	205 +	6	4
400	1·0018736	18785 +	157	203	6	4

TABLE II.

Values of the 12th Incomplete Normal Moment Function.

x	$m_{12}(x)$	δ^2	x	$m_{12}(x)$	δ^2
0·0	·0000000		2·5	·0317107	+23061
0·1	·0000000		2·6	·0429539	25057
0·2	·0000000		2·7	·0567028	26157
0·3	·0000000		2·8	·0730674	26170
0·4	·0000000		2·9	·0920490	24977
0·5	·0000000		3·0	·1135283	22562
0·6	·0000000		3·1	·1372638	19005
0·7	·0000000		3·2	·1628998	14482
0·8	·0000001	+ 4	3·3	·1899840	9265
0·9	·0000006	7	3·4	·2179947	3657
1·0	·0000018	30	3·5	·2463711	− 2004
1·1	·0000060	68	3·6	·2745471	7391
1·2	·0000170	152	3·7	·3019840	12221
1·3	·0000432	315	3·8	·3281988	16270
1·4	·0001009	601	3·9	·3527866	19393
1·5	·0002187	1072	4·0	·3754351	21516
1·6	·0004437	1799	4·1	·3959320	22653
1·7	·0008486	2860	4·2	·4141636	22875
1·8	·0015395 −	4312	4·3	·4301077	22306
1·9	·0026616	6206	4·4	·4438212	21095
2·0	·0044043	8544	4·5	·4554252	19410
2·1	·0070014	11274	4·6	·4650882	17426
2·2	·0107259	14287	4·7	·4730086	15284
2·3	·0158791	17413	4·8	·4794006	13117
2·4	·0227736	20426	4·9	·4844809	11030
			5·0	·4884582	9097
			∞	·5000000	

ON THE MOMENTS OF THE DISTRIBUTION OF SQUARED STANDARD-DEVIATIONS FOR SAMPLES OF *N* DRAWN FROM AN INDEFINITELY LARGE POPULATION.

By A. E. R. CHURCH, M.A., B.Sc.

(i) Being engaged in an experimental investigation of the moments of the distributions of means and squared standard-deviations (σ^2) of samples of N drawn from a definitely skew population I have had occasion to resort to the formulae for the moments of the distribution of σ^2 for samples of N drawn from an indefinitely large population given by Prof. Al. A. Tchouproff in *Biometrika*, Vol. XII. p. 193 (23) and p. 194 (24) and (25).

Considerable time was spent in an attempt to apply these to the actual distributions used, and finally Prof. Karl Pearson, under whose direction this experimental work is being carried out, came to the definite conclusion that the formula (25) for the 4th moment was in error. I read through the algebraic manipulation which leads up to these numerical results but was unable to follow its final stages where the statement is very exiguous.

Now, previously to Tchouproff, "Student" had given the first and second of these moments in *Biometrika*, Vol. VI. p. 3, although unfortunately, the second moment contained by a printer's failure slight errors corrected, however, in the following line. "Student" also gave the 3rd and 4th moments by the same process in the case of samples from a normal distribution..

In the present paper I have extended "Student's" method to obtain the 3rd and 4th moments for samples from any population so that I might have a revision of Tchouproff's 4th moment by an independent method. I further desired to curtail, if possible, the somewhat extensive use of symbols employed in Prof. Tchouproff's memoir, which discourages the English statistical student.

(ii) Let $\bar{\nu}_t$ be the tth moment of the original indefinitely large population of a variable X, where X is measured from the mean of this population.

Let ${}_2M_t'$ be the tth moment, about $X = 0$, of the distribution of σ^3 of samples of N drawn from this population.

Let ${}_2M_t$ be the tth moment, about its own mean, of this distribution of σ^2.

Then

$${}_2M_t' = \Sigma \left[\frac{S(X_1^2)}{N} - \left\{ \frac{S(X_1)}{N} \right. \right.$$

where S refers to the summation over the N values $X_1, X_2, \dots X_N$ of the variable X which form one sample, and where, since the original population is indefinitely large, these N values of X may be considered independent or uncorrelated; while by Σ is meant the process of summing for all possible samples of N and dividing by the number of such samples.

To obtain the first four moments ${}_2M_1'$, ${}_2M_2'$, ${}_2M_3'$, ${}_2M_4'$ it is necessary to expand a number of expressions of the form $\{S(X_1)^2\}^p\,\{S(X_1)\}^{2q}$, where p and q are zero or positive integers less than 8, in series of sums like $S(X_1^{\alpha}X_2^{\beta}\ldots X_r^{\lambda})$, where $\alpha, \beta, \ldots \lambda$ and r are positive integers or zero, each series of sums being isobaric in the X's of weight m, where $m = 2, 4, 6$, or 8.

Thus I find

$$\{S(X_1^2)\}^2 = S(X_1^4) + 2\,S(X_1^2X_2^2),$$

$$\{S(X_1^2)\}^3 = S(X_1^6) + 3\,S(X_1^4X_2^2) + 6\,S(X_1^2X_2^2X_3^2),$$

$$\begin{aligned}\{S(X_1^2)\}^4 = {} & S(X_1^8) + 4\,S(X_1^6X_2^2) + 12\,S(X_1^4X_2^2X_3^2) + 6\,S(X_1^4X_2^4)\\ & + 24\,S(X_1^2X_2^2X_3^2X_4^2),\end{aligned}$$

$$\{S(X_1)\}^2 = S(X_1^2) + 2\,S(X_1X_2),$$

$$\begin{aligned}\{S(X_1)\}^4 = {} & S(X_1^4) + 4\,S(X_1^3X_2) + 6\,S(X_1^2X_2^2) + 12\,S(X_1^2X_2X_3)\\ & + 24\,S(X_1X_2X_3X_4),\end{aligned}$$

$$\begin{aligned}\{S(X_1)\}^6 = {} & S(X_1^6) + 6\,S(X_1^5X_2) + 15\,S(X_1^4X_2^2) + 30\,S(X_1^4X_2X_3)\\ & + 20\,S(X_1^3X_2^3) + 60\,S(X_1^3X_2^2X_3) + 120\,S(X_1^3X_2X_3X_4)\\ & + 90\,S(X_1^2X_2^2X_3^2) + 180\,S(X_1^2X_2^2X_3X_4) + 360\,S(X_1^2X_2X_3X_4X_5)\\ & + 720\,S(X_1X_2X_3X_4X_5X_6),\end{aligned}$$

$$\begin{aligned}\{S(X_1)\}^8 = {} & S(X_1^8) + 8\,S(X_1^7X_2) + 28\,S(X_1^6X_2^2) + 56\,S(X_1^6X_2X_3)\\ & + 56\,S(X_1^5X_2^3) + 168\,S(X_1^5X_2^2X_3) + 336\,S(X_1^5X_2X_3X_4)\\ & + 70\,S(X_1^4X_2^4) + 280\,S(X_1^4X_2^3X_3) + 420\,S(X_1^4X_2^2X_3^2)\\ & + 840\,S(X_1^4X_2^2X_3X_4) + 1680\,S(X_1^4X_2X_3X_4X_5)\\ & + 560\,S(X_1^3X_2^3X_3^2) + 1120\,S(X_1^3X_2^3X_3X_4) + 1680\,S(X_1^3X_2^2X_3^2X_4)\\ & + 3360\,S(X_1^3X_2^2X_3X_4X_5) + 6720\,S(X_1^3X_2X_3X_4X_5X_6)\\ & + 2520\,S(X_1^2X_2^2X_3^2X_4^2) + 5040\,S(X_1^2X_2^2X_3^2X_4X_5)\\ & + 10080\,S(X_1^2X_2^2X_3X_4X_5X_6) + 20160\,S(X_1^2X_2X_3X_4X_5X_6X_7)\\ & + 40320\,S(X_1X_2X_3X_4X_5X_6X_7X_8),\end{aligned}$$

$$\begin{aligned}\{S(X_1^2)\}^2\{S(X_1)\}^2 = {} & S(X_1^6) + 2\,S(X_1^5X_2) + 3\,S(X_1^4X_2^2) + 2\,S(X_1^4X_2X_3)\\ & + 4\,S(X_1^3X_2^3) + 4\,S(X_1^3X_2^2X_3) + 6\,S(X_1^2X_2^2X_3^2)\\ & + 4\,S(X_1^2X_2^2X_3X_4),\end{aligned}$$

$$\begin{aligned}\{S(X_1^2)\}^2\{S(X_1)\}^4 = {} & S(X_1^8) + 4\,S(X_1^7X_2) + 8\,S(X_1^6X_2^2) + 12\,S(X_1^6X_2X_3)\\ & + 12\,S(X_1^5X_2^3) + 20\,S(X_1^5X_2^2X_3) + 24\,S(X_1^5X_2X_3X_4)\\ & + 14\,S(X_1^4X_2^4) + 32\,S(X_1^4X_2^2X_3^2) + 28\,S(X_1^4X_2^3X_3)\\ & + 36\,S(X_1^4X_2^2X_3X_4) + 24\,S(X_1^4X_2X_3X_4X_5) + 40\,S(X_1^3X_2^3X_3^2)\\ & + 56\,S(X_1^3X_2^2X_3^2X_4) + 48\,S(X_1^3X_2^3X_3X_4) + 48\,S(X_1^3X_2^2X_3X_4X_5)\\ & + 72\,S(X_1^2X_2^2X_3^2X_4^2) + 72\,S(X_1^2X_2^2X_3^2X_4X_5)\\ & + 48\,S(X_1^2X_2^2X_3X_4X_5X_6),\end{aligned}$$

$$\begin{aligned}\{S(X_1^2)\}^3\{S(X_1)\}^2 = {} & S(X_1^8) + 2\,S(X_1^7X_2) + 4\,S(X_1^6X_2^2) + 2\,S(X_1^6X_2X_3)\\ & + 6\,S(X_1^5X_2^3) + 6\,S(X_1^5X_2^2X_3) + 6\,S(X_1^4X_2^4) + 6\,S(X_1^4X_2^3X_3)\\ & + 12\,S(X_1^4X_2^2X_3^2) + 6\,S(X_1^4X_2^2X_3X_4) + 12\,S(X_1^3X_2^3X_3^2)\\ & + 12\,S(X_1^3X_2^2X_3^2X_4) + 24\,S(X_1^2X_2^2X_3^2X_4^2)\\ & + 12\,S(X_1^2X_2^2X_3^2X_4X_5),\end{aligned}$$

$$\{S(X_1^2)\}\{S(X_1)\}^2 = S(X_1^4) + 2S(X_1^3X_2) + 2S(X_1^2X_2^2) + 2S(X_1^2X_2X_3),$$

$$\begin{aligned}\{S(X_1^2)\}\{S(X_1)\}^4 = {} & S(X_1^6) + 4S(X_1^5X_2) + 7S(X_1^4X_2^2) + 12S(X_1^4X_2X_3)\\ & + 8S(X_1^3X_2^3) + 16S(X_1^3X_2^2X_3) + 24S(X_1^3X_2X_3X_4)\\ & + 18S(X_1^2X_2^2X_3^2) + 24S(X_1^2X_2^2X_3X_4) + 24S(X_1^2X_2X_3X_4X_5),\end{aligned}$$

$$\begin{aligned}\{S(X_1^2)\}\{S(X_1)\}^6 = {} & S(X_1^8) + 6S(X_1^7X_2) + 16S(X_1^6X_2^2) + 30S(X_1^6X_2X_3)\\ & + 26S(X_1^5X_2^3) + 66S(X_1^5X_2^2X_3) + 120S(X_1^5X_2X_3X_4)\\ & + 30S(X_1^4X_2^4) + 90S(X_1^4X_2^3X_3) + 120S(X_1^4X_2^2X_3^2)\\ & + 210S(X_1^4X_2^2X_3X_4) + 360S(X_1^4X_2X_3X_4X_5) + 140S(X_1^3X_2^3X_3^2)\\ & + 300S(X_1^3X_2^2X_3^2X_4) + 240S(X_1^3X_2^3X_3X_4)\\ & + 480S(X_1^3X_2^2X_3X_4X_5) + 720S(X_1^3X_2X_3X_4Y_5X_6)\\ & + 360S(X_1^2X_2^2X_3^2X_4^2) + 540S(X_1^2X_2^2X_3^2X_4X_5)\\ & + 720S(X_1^2X_2^2X_3X_4X_5X_6) + 720S(X_1^2X_2X_3X_4X_5X_6X_7).\end{aligned}$$

Now if $\alpha, \beta, \gamma, \dots$ be all different, $S(X_1^\alpha X_2^\beta \dots X_r^\lambda)$ has $N(N-1)\dots(N-r+1)$ terms in its expansion; but if p indices be alike of one kind α, and q be alike of another kind β, etc., it has $\dfrac{N(N-1)\dots(N-r+1)}{p!\,q!\dots}$ terms in its expansion:

$$\therefore\ \Sigma S(X_1^\alpha X_2^\beta \dots X_r^\lambda) = N(N-1)\dots(N-r+1)\,\Sigma'(X_1^\alpha X_2^\beta \dots X_r^\lambda),$$

or
$$\frac{N(N-1)\dots(N-r+1)}{p!\,q!\dots}\,\Sigma'\{X_1^\alpha X_2^\alpha \dots X_p^\alpha X^\beta_{p+1} \dots X^\beta_{p+q} \dots\},$$

as the case may be, where Σ' refers to the summation over all values of the variate X in the original population, divided by the number of such values of course.

Now since $X_1, X_2, \dots X_N$ are independent or uncorrelated we have therefore

$$\Sigma S(X_1^\alpha X_2^\beta \dots X_r^\lambda) = N(N-1)\dots(N-r+1)\,\Sigma' X_1^\alpha\, \Sigma' X_2^\beta \dots \Sigma' X_r^\lambda$$

or
$$\frac{N(N-1)\dots(N-r+1)}{p!\,q!\dots}(\Sigma' X_1^\alpha)^p(\Sigma' X_2^\beta)^q \dots,$$

that is
$$\Sigma S(X_1^\alpha X_2^\beta \dots X_r^\lambda) = N(N-1)\dots(N-r+1)\,\bar{\nu}_\alpha \bar{\nu}_\beta \dots \bar{\nu}_\lambda,$$

or
$$\frac{N(N-1)\dots(N-r+1)}{p!\,q!\dots}\,\bar{\nu}_\alpha^{\,p}\bar{\nu}_\beta^{\,q}\dots;$$

but since these moments of the original distribution are about its mean $X = 0$, we have $\bar{\nu}_1 = 0$, therefore if any of $\alpha, \beta, \dots \lambda$ are unity, $\Sigma S(X_1^\alpha X_2^\beta \dots X_r^\lambda) = 0$.

Therefore, using the above, we have

$${}_2M_1' = \Sigma\left[\frac{S(X_1^2)}{N} - \frac{\{S(X_1)\}^2}{N^2}\right] = \bar{\nu}_2\left(1 - \frac{1}{N}\right),$$

also
$${}_2M_2' = \Sigma\left[\frac{\{S(X_1^2)\}^2}{N^2} - \frac{2\{S(X_1^2)\}\{S(X_1)\}^2}{N^3} + \frac{\{S(X_1)\}^4}{N^4}\right],$$

i.e.
$${}_2M_2' = \left(1 - \frac{1}{N}\right)\left[\frac{1}{N}\left(1 - \frac{1}{N}\right)\bar{\nu}_4 + \left(1 - \frac{2}{N} + \frac{3}{N^2}\right)\bar{\nu}_2^2\right].$$

Further

$${}_2M_3' = \Sigma\left[\frac{\{S(X_1^2)\}^3}{N^3} - \frac{3\{S(X_1^2)\}^2\{S(X_1)\}^2}{N^4} + \frac{3\{S(X_1^2)\}\{S(X_1)\}^4}{N^5} - \frac{\{S(X_1)\}^6}{N^6}\right],$$

i.e.

$$\begin{aligned}{}_2M_3' = {} & \bar{\nu}_2^3 + \frac{1}{N}[3\bar{\nu}_4\bar{\nu}_2 - 6\bar{\nu}_2^3] + \frac{1}{N^2}[\bar{\nu}_6 - 12\bar{\nu}_4\bar{\nu}_2 - 6\bar{\nu}_3^2 + 20\bar{\nu}_2^3] \\ & - \frac{1}{N^3}[3\bar{\nu}_6 - 30\bar{\nu}_4\bar{\nu}_2 - 18\bar{\nu}_3^2 + 48\bar{\nu}_2^3] + \frac{1}{N^4}[3\bar{\nu}_6 - 36\bar{\nu}_4\bar{\nu}_2 - 22\bar{\nu}_3^2 + 63\bar{\nu}_2^3] \\ & - \frac{1}{N^5}[\bar{\nu}_6 - 15\bar{\nu}_4\bar{\nu}_2 - 10\bar{\nu}_3^2 + 30\bar{\nu}_2^3].\end{aligned}$$

Again

$$\begin{aligned}{}_2M_4' = \Sigma\Bigg[& \frac{\{S(X_1^2)\}^4}{N^4} - \frac{4\{S(X_1^2)\}^3\{S(X_1)\}^2}{N^5} + \frac{6\{S(X_1^2)\}^2\{S(X_1)\}^4}{N^6} \\ & - \frac{4\{S(X_1^2)\}\{S(X_1)\}^6}{N^7} + \frac{\{S(X_1)\}^8}{N^8}\Bigg],\end{aligned}$$

i.e.

$$\begin{aligned}{}_2M_4' = {} & \bar{\nu}_2^4 + \frac{1}{N}[6\bar{\nu}_4\bar{\nu}_2^2 - 10\bar{\nu}_2^4] + \frac{1}{N^2}[4\bar{\nu}_6\bar{\nu}_2 + 3\bar{\nu}_4^2 - 42\bar{\nu}_4\bar{\nu}_2^2 - 24\bar{\nu}_3^2\bar{\nu}_2 + 53\bar{\nu}_2^4] \\ & + \frac{1}{N^3}[\bar{\nu}_8 - 20\bar{\nu}_6\bar{\nu}_2 - 24\bar{\nu}_5\bar{\nu}_3 - 15\bar{\nu}_4^2 + 180\bar{\nu}_4\bar{\nu}_2^2 + 192\bar{\nu}_3^2\bar{\nu}_2 - 218\bar{\nu}_2^4] \\ & - \frac{1}{N^4}[4\bar{\nu}_8 - 64\bar{\nu}_6\bar{\nu}_2 - 96\bar{\nu}_5\bar{\nu}_3 - 54\bar{\nu}_4^2 + 576\bar{\nu}_4\bar{\nu}_2^2 + 688\bar{\nu}_3^2\bar{\nu}_2 - 687\bar{\nu}_2^4] \\ & + \frac{1}{N^5}[6\bar{\nu}_8 - 112\bar{\nu}_6\bar{\nu}_2 - 176\bar{\nu}_5\bar{\nu}_3 - 102\bar{\nu}_4^2 + 1122\bar{\nu}_4\bar{\nu}_2^2 + 1360\bar{\nu}_3^2\bar{\nu}_2 - 1398\bar{\nu}_2^4] \\ & - \frac{1}{N^6}[4\bar{\nu}_8 - 92\bar{\nu}_6\bar{\nu}_2 - 160\bar{\nu}_5\bar{\nu}_3 - 95\bar{\nu}_4^2 + 1110\bar{\nu}_4\bar{\nu}_2^2 + 1400\bar{\nu}_3^2\bar{\nu}_2 - 1515\bar{\nu}_2^4] \\ & + \frac{1}{N^7}[\bar{\nu}_8 - 28\bar{\nu}_6\bar{\nu}_2 - 56\bar{\nu}_5\bar{\nu}_3 - 35\bar{\nu}_4^2 + 420\bar{\nu}_4\bar{\nu}_2^2 + 560\bar{\nu}_3^2\bar{\nu}_2 - 630\bar{\nu}_2^4].\end{aligned}$$

Now changing the origin from the mean of the original population to the mean of this distribution of σ^2 of samples of N we have for the equations of transformation the usual formulae:

$$\begin{aligned}{}_2M_2 &= {}_2M_2' - ({}_2M_1')^2, \\ {}_2M_3 &= {}_2M_3' - 3\,.\,{}_2M_2\,.\,{}_2M_1' - ({}_2M_1')^3, \\ {}_2M_4 &= {}_2M_4' - 4\,.\,{}_2M_3\,.\,{}_2M_1' - 6\,.\,{}_2M_2\,.\,({}_2M_1')^2 - ({}_2M_1')^4,\end{aligned}$$

which give

$$\begin{aligned}{}_2M_2 &= \frac{1}{N}\left(1-\frac{1}{N}\right)\left[\left(1-\frac{1}{N}\right)\bar{\nu}_4 - \left(1-\frac{3}{N}\right)\bar{\nu}_2^2\right] \\ &= \frac{1}{N}[\bar{\nu}_4 - \bar{\nu}_2^2] - \frac{2}{N^2}[\bar{\nu}_4 - 2\bar{\nu}_2^2] + \frac{1}{N^3}[\bar{\nu}_4 - 3\bar{\nu}_2^2] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(i)},\end{aligned}$$

$$\begin{aligned}{}_2M_3 = {} & \frac{1}{N^2}[\bar{\nu}_6 - 3\bar{\nu}_4\bar{\nu}_2 - 6\bar{\nu}_3^2 + 2\bar{\nu}_2^3] - \frac{1}{N^3}[3\bar{\nu}_6 - 21\bar{\nu}_4\bar{\nu}_2 - 18\bar{\nu}_3^2 + 26\bar{\nu}_2^3] \\ & + \frac{1}{N^4}[3\bar{\nu}_6 - 33\bar{\nu}_4\bar{\nu}_2 - 22\bar{\nu}_3^2 + 54\bar{\nu}_2^3] - \frac{1}{N^5}[\bar{\nu}_6 - 15\bar{\nu}_4\bar{\nu}_2 - 10\bar{\nu}_3^2 + 30\bar{\nu}_2^3] \quad \text{(ii)},\end{aligned}$$

and

$$
\begin{aligned}
{}_2M_4 = {} & \frac{3}{N^2}[\bar{\nu}_4 - \bar{\nu}_2^2]^2 + \frac{1}{N^3}[\bar{\nu}_8 - 4\bar{\nu}_6\bar{\nu}_2 - 24\bar{\nu}_5\bar{\nu}_3 - 15\bar{\nu}_4^2 + 48\bar{\nu}_4\bar{\nu}_2^2 + 96\bar{\nu}_3^2\bar{\nu}_2 - 30\bar{\nu}_2^4] \\
& - \frac{1}{N^4}[4\bar{\nu}_8 - 40\bar{\nu}_6\bar{\nu}_2 - 96\bar{\nu}_5\bar{\nu}_3 - 54\bar{\nu}_4^2 + 336\bar{\nu}_4\bar{\nu}_2^2 + 528\bar{\nu}_3^2\bar{\nu}_2 - 306\bar{\nu}_2^4] \\
& + \frac{1}{N^5}[6\bar{\nu}_8 - 96\bar{\nu}_6\bar{\nu}_2 - 176\bar{\nu}_5\bar{\nu}_3 - 102\bar{\nu}_4^2 + 924\bar{\nu}_4\bar{\nu}_2^2 + 1232\bar{\nu}_3^2\bar{\nu}_2 - 1044\bar{\nu}_2^4] \\
& - \frac{1}{N^6}[4\bar{\nu}_8 - 88\bar{\nu}_6\bar{\nu}_2 - 160\bar{\nu}_5\bar{\nu}_3 - 95\bar{\nu}_4^2 + 1050\bar{\nu}_4\bar{\nu}_2^2 + 1360\bar{\nu}_3^2\bar{\nu}_2 - 1395\bar{\nu}_2^4] \\
& + \frac{1}{N^7}[\bar{\nu}_8 - 28\bar{\nu}_6\bar{\nu}_2 - 56\bar{\nu}_5\bar{\nu}_3 - 35\bar{\nu}_4^2 + 420\bar{\nu}_4\bar{\nu}_2^2 + 560\bar{\nu}_3^2\bar{\nu}_2 - 630\bar{\nu}_2^4] \quad \text{...(iii).}
\end{aligned}
$$

A comparison between the formulae given above for ${}_2M_4'$ and ${}_2M_4$ with Tchouproff's formula (25) appears to indicate some algebraical slip in his substitution of the value of his $W^{(4)}_{2(N)}$ of equation (21) of p. 193 in the third equation of his (10) p. 187 to get his (25) of p. 194. For his coefficients in the terms of (25) involving $1/N^6$ are exactly those of $1/N^2$ in his $W^{(4)}_{2(N)}$, or of the term in $1/N^6$ in my ${}_2M_4'$, and not those of this fourth moment, our ${}_2M_4$, when transferred to the mean σ^2 of samples.

The method, due to "Student," employed here is undoubtedly more direct and concise in statement than that employed by Prof. Tchouproff; further it presents fewer algebraical difficulties and the algebra has the advantage of being easily followed and checked at each stage.

It is difficult to estimate at present the real value of these high moment formulae in actual practice, but it is fairly clear that, if they be used to predict the results of any actual sampling, that sampling must be strictly in accordance with the assumptions involved and must be very carefully carried out in experimental work before any reasonable agreement can be obtained.

In conclusion I must express my thanks to Prof. K. Pearson for his continual kindly advice and criticism.

THE CORPUSCLE PROBLEM.

A MATHEMATICAL STUDY OF A BIOMETRIC PROBLEM.

By S. D. WICKSELL, *Lund*.

1. IN several organs of the human body, as well as that of many animals, a microscopic and sometimes also a macroscopic inspection reveals a number of small, more or less regularly formed corpuscles embedded in the tissue of the organ. Such corpuscles are for instance the Germ Centres or secondary follicles (Hellman) found in lymphoidal organs—as the lymphatic glands, the tonsils or the spleen,—the so-called *Hassal corpuscles* in the thymus, the *Langerhans insulae* in the pancreas, etc. Generally these bodies vary considerably in size within the same organ, and in different organs or different individuals the number, as well as the size distribution of the corpuscles will be found to be different.

In anatomy as in other biological disciplines the last few decades have seen an enormous increase in the demand for exact and detailed information. The modern anatomist often devotes the labour and toil of many years to the study of one special organ, patiently collecting and measuring extensive material from animals or *post mortem* subjects. The more extensive the material the more difficult it will be, however, to perform for each individual all the detailed observations required; the investigator will have to look for simplifications and short-cut methods of observation, as far as this is possible without making the resulting information insufficient. In studying the number and size distributions of the above-mentioned corpuscles, the anatomists have, then, run up against a mathematical problem, which is, even in the simpler cases, beyond the powers of persons unacquainted with higher mathematics. This problem, which I call the "corpuscle problem," is, as far as I know, as yet unsolved, although several anatomists have tried to reach a solution along more or less purely empirical lines[*]. At the request of Professor T. Hellman, of Lund, I have therefore taken the question up from the mathematico-statistical point of view. My interest in the matter has subsequently been greatly increased by the rather curious fact, that the problem is in essential features identical with an important problem in astronomical statistics.

2. In order to make the matter more concrete I shall begin with a short description of the observations performed by Hellman regarding the so-called secondary follicles (Germ Centres) in the lymphatic tissue of the spleen.

---

[*] See for instance J. A. Hammar, "Methode, die Menge der Rinde und des Marks des Thymus, sowie die Anzahl und die Grösse der Hassal'schen Körper zahlenmässig festzustellen" (*Zeitschrift f. angewandte Anatomie und Konstitutionslehre*, Bd. I, Heft 4/5, 1914); J. A. Hammar and T. Hellman, "Ein Fall von Tyreoaplasie unter Berücksichtigung der innersekretorischen und lymphoiden Organe: Abteilung II: Die lymphoiden Organe," von T. Hellman (*Zeitschrift f. die gesammte Anatomie*, Zweite Abteilung, Bd. VIII, Heft 4,

The Germ Centre, or rather the secondary follicle in the Malpighian corpuscle, is a small, approximately spherical, body embedded in the tissue of the spleen. The number of such corpuscles in different spleens is very variable, ranging from a very few or none at all up to hundreds of thousands per organ. The sizes are also very variable, the same organ often containing all sizes from a small fraction of a millimetre up to one millimetre in diameter.

The observations are arranged in the following way: The weight of each spleen being ascertained, and fixation and preparation being properly attended to, microtome sections of 0·018 mm. in thickness are taken from different parts of the organ. By the aid of a projection apparatus a picture of each section is obtained in 18-fold magnification on a paper, and the contours (generally circular, sometimes oval) of the follicles are traced on the paper in lead pencil. On the drawings thus obtained the diameter of every corpuscle contour is measured to the nearest millimetre; when the image is not circular the largest diameter and also the diameter at right angles to the largest are measured, the arithmetic mean of the two measures being taken as value of the diameter. When the diameter is below 1·5 mm. in this magnification the measures are rejected on account of difficulties of identification. From the protocols kept of these measures the distribution of the diameters of the images in the section is obtained for classes of one mm. breadth (1·5—2·5, 2·5—3·5, etc.).

The outline of the preparation also being marked out in the drawing the parts of the paper representing the spleen parenchyma may be cut out as well as the images of the follicles. As a carefully selected homogeneous paper is always used the areas of the sections may be determined by weighing the drawings. The weights of the drawings before and after the images of the follicles have been cut away will also give sample values of the ratio of the total volume of the follicles to the total volume of the spleen. As will be seen later on it is necessary to know either the area of the section or this ratio in order to determine the total number of follicles in the organ. Of course, when several sections have been taken and measured from different parts of the same organ, they are put together and regarded as one single section of correspondingly greater area. In the following we will regard this area as great enough for the purposes of generalization. Thus the question of the errors of sampling will not be taken up in this connection.

It is furthermore clear that the diameters of the images in the drawings are generally not the true diameters of the follicles. Only when the section chances to pass through the centre of a follicle will this be the case. As a rule the centres of the follicles will lie more or less distant from the plane of the section. The more distant the centre the smaller will the image diameter be, in comparison to the actual diameter. As the distance of the centre is generally unknown we see that the diameters measured in the drawings are "apparent diameters" in the same sense as the projection on the heavens of the motion of a star or the distance of the components of a double star is called apparent, the inclination to the plane of projection of the true motion or line joining the components being unknown. Indeed, the difficulty of the anatomist at his microscope is essentially the same as the difficulty of the astronomer at his telescope; he only sees the projection of things. The statistical problem met with is also in both cases similar. The "true" character being a function of the apparent character and an unknown variable, and making a plausible assumption as to the distribution of this variable, the problem is to express the distribution of the true character in terms of the observed distribution of the apparent character. It will be seen later on that there is also one problem in stellar statistics, which is in essential details identical to the anatomic corpuscle problem, namely the problem of determining the distribution in space of the stars in a stellar cluster, from the projected distribution on the heavens. Hence some of the methods here worked out for anatomic applications may be of value also for the astronomers.

Evidently the problem arising in connection with the secondary follicles of the spleen is the following: To find the distribution of the actual ("true") diameters of the follicles from the distribution of the diameters of the images (the "apparent" diameters) in the microtome sections. That these two distributions are generally different is clear from the following two reasons: (1) The apparent diameters are as a rule smaller than the true diameters, and (2) a random section will contain a relatively greater number of actually large follicles, than of small ones, because the former more frequently will come within reach of the section plane. The two causes will, however, work in opposite directions, thus having a chance to balance each other. It will be seen in the following that under certain conditions an exact balance takes place, making the relative distributions of the true and the apparent diameters equal.

We must mention finally that in taking a series of successive sections instead of only one or a few sample sections, the true diameters may be directly observed. In this respect the position of the microscopist is more favourable than that of his colleague at the telescope. But as the measuring of one single section is generally very laborious, and as the material studied embraces several hundred spleens, it has been found impossible to use this method, except in special cases to obtain a control and empirical test of the theoretical reductions. These reductions naturally involve the assumption that the corpuscles are approximately spherical in form, and that they are fairly uniformly distributed throughout the whole organ. Series of successive sections actually measured show these conditions to be sufficiently fulfilled in the case of the secondary follicles of the spleen. Attempts have been made to use the serial sections for obtaining a set of purely empirical equations of reduction. These trials have, however, shown that the theoretical equations are more to be relied upon.

I now pass on to a more general formulation and solution of the "corpuscle problem."

3. *The Corpuscle Problem.* In an opaque body there are suspended a large number of spherical corpuscles of different sizes, the density of the centres and the distribution of the sizes being the same in all parts of the body. This body is split in two by a plane section. To express the distribution of the diameters of the

corpuscles in terms of the distribution of the diameters of the circular contours found in the plane section.

Solution. Let the number of centres per unit volume be N and let $F(r)\,dr$ be the relative frequency of corpuscles with the diameter in the interval r to $r+dr$. We now have

$$\int_0^R F(r)\,dr = 1 \quad \text{.................................(1).}$$

R being the upper limit to the diameters. Furthermore the number of corpuscles of diameter r to $r+dr$ coming within reach of an arbitrary plane is

$$NrF(r)\,dr$$

per unit area in the plane. Hence, r_0 being the mean diameter of the corpuscles· Nr_0 will be the total number of corpuscle images to the unit area in the section plane. Putting

$$Nr_0 f(r) = NrF(r) \quad \text{..............................(2).}$$

it is seen that $f(r)\,dr$ is the relative frequency of corpuscles cut by the plane. having the actual diameter between r and $r+dr$. We thus have

$$f(r) = \frac{rF(r)}{r_0} \quad \text{....................................(3).}$$

The probability that a corpuscle with the diameter r, falling within the reach of the plane section, has its centre at the distance y to $y+dy$ from this plane, is evidently equal to

$$\frac{2dy}{r}.$$

Hence the relative number of corpuscles of diameter r to $r+dr$ and with their centres at the distance y to $y+dy$ $(y<r)$ from the plane, is equal to

$$\frac{2}{r}f(r)\,dy\,dr.$$

But the "apparent" diameter of the circular image cut out of a corpuscle by the plane being denoted by x, we have

$$x^2 = r^2 - 4y^2,$$

and it is seen that the relative number of corpuscles having the actual diameter r to $r+dr$ and the apparent diameter x to $x+dx$ in the section plane is

$$\frac{2}{r}f(r)\,dr\frac{dy}{dx}dx = \frac{1}{r}f(r)\,dr\frac{x\,dx}{\sqrt{r^2-x^2}} = \frac{1}{r_0}F(r)\,dr\frac{x\,dx}{\sqrt{r^2-x^2}} \quad \text{...........(4).}$$

Hence, $\phi(x)$ being the relative frequency function of the apparent diameters in the plane section, we have

$$\phi(x) = \frac{x}{r_0}\int_x^R F(r)\frac{dr}{\sqrt{r^2-x^2}} \quad \text{..........................(5).}$$

As $\phi(x)$ is here the function supposed to be known, this is an integral equation to determine the sought function $F(r)$.

This integral equation may easily be transformed to the same form as the well-known integral equation of Abel*, and thus the solution is readily obtained. It will be still more convenient, however, to apply the solution given by von Zeipel of the following equation, occurring in the reduction of the observations of globular clusters †,

$$\psi(x) = 2\int_x^R \chi(r)\frac{r\,dr}{\sqrt{r^2-x^2}} \quad \text{..........................(6).}$$

For this equation von Zeipel has obtained the solution

$$\chi(r) = -\frac{1}{\pi}\int_r^R \frac{\psi'(x)\,dx}{\sqrt{x^2-r^2}} = \frac{1}{\pi}\int_0^{\sqrt{R^2-r^2}} P(\sqrt{l^2+r^2})\,dl \quad \text{...........(7),}$$

where
$$P(x) = -\frac{1}{x}\frac{d\psi(x)}{dx}.$$

The solution of (5) consequently assumes the attractive form

$$F(r) = -\frac{2rr_0}{\pi}\int_r^R \frac{d\left(\frac{\phi(x)}{x}\right)}{dx}\frac{dx}{\sqrt{x^2-r^2}} = \frac{2rr_0}{\pi}\int_0^{\sqrt{R^2-r^2}} P(\sqrt{l^2+r^2})\,dl \quad \text{...(8),}$$

where
$$P(x) = -\frac{1}{x}\frac{d\left(\frac{\phi(x)}{x}\right)}{dx} \quad \text{..............................(9).}$$

4. This form of solution is also specially adapted to the case, in which we do not know an analytical expression for $\phi(x)$, but only have a series of numerical values of this function, which is what takes place both in the astronomical problem as well as in the anatomical problem here in question.

The following method will certainly be sufficiently accurate for the anatomical applications: The observations in the sections having been arranged in classes, the relative frequency-histogram of the apparent diameters may be constructed. Smoothing this by a continuous curve (a free-hand curve will suffice) we may read off the ordinates $\phi(i)$; $(i = 0, 1, 2, \ldots R)$. Thence we obtain the series $\phi(1):1$; $\phi(2):2$; $\phi(3):3$, etc., and finally the differential quotients of $\phi(x):x$ may be obtained by the aid of the well-known central difference formula

$$f'(x) = \Delta_0^1 f(x) - \tfrac{1}{6}\Delta_0^3 f(x) + \tfrac{1}{30}\Delta_0^5 f(x) + \ldots.$$

From these we easily obtain the values of $P(i)$, and the curve $y = P(x)$ may be plotted, so that we may read off the values of

$$P(\sqrt{i^2+r^2});\quad (i = 0, 1, 2 \ldots R).$$

---

* See for instance Bôcher, *An Introduction to the Study of Integral Equations*, pp. 6—11 (Cambridge Tracts in Mathematics and Mathematical Physics), 1909.

† H. von Zeipel, "Catalogue de l'amas globulaire Messier 3" (*Annales de l'Observatoire de Paris*, Tome XXV, 1906, p. 80); "Recherches sur la constitution des amas globulaires" (*Kungl. Svenska Vetenskapsakademiens Handlingar*, Bd. LI, No. 5, 1913, p. 9). H. von Zeipel und J. Lindgren, "Photometrische Untersuchungen der Sterngruppe Messier 37" (*Kungl. Svenska Vetenskapsakademiens Handlingar*, Bd. LXI, No. 15, 1921, pp. 109—110).

Using the approximate quadrature formula *

$$\frac{2r}{\pi}\int_0^{\sqrt{R^2-r^2}} P(\sqrt{l^2+r^2})\,dl = \frac{2r}{\pi}\left[\sum_{i=0}^{\sqrt{R^2-r^2}} P(\sqrt{i^2+r^2}) - \tfrac{1}{2}P(r)\right] = A(r) \quad \ldots(10),$$

we have

$$F(r) = r_0 A(r).$$

Since

$$\int_0^R F(r)\,dr = 1,$$

we evidently must have approximately $\frac{1}{r_0} = \sum_{r=0}^{R} A(r)$. The mean diameter r_0 being found in this manner, a graph of the sought function $F(r)$ is easily constructed.

To determine the total number N of follicles per unit volume we may use the results obtained by weighing the section-drawings. These weighings give, as mentioned in § 2, the volume of the follicles in percentages of the spleen tissue. Denoting this by 100 V, and the mean volume of the follicles by v_0, we evidently have

$$N = V : v_0.$$

The mean volume v_0 is obtained from the equation

$$v_0 = \frac{\pi}{6}\sum_{r=0}^{R} r^3 F(r).$$

The weighings of the drawings may, however, also be used to determine the area of the observed sections. This area being denoted by Y, and the number of follicles represented in the observed sections being denoted by n, we evidently have

$$n = N r_0 Y,$$

from which equation N is readily found. It may be presumed that this method is subject to a smaller mean error than the preceding method.

5. Another solution of (5), which is perhaps more to the taste of the trained statistician, is obtained by the aid of the method of moments.

The quantities v_n' and N_n', defined by the equations

$$v_n' = \int_0^R x^n \phi(x)\,dx; \quad N_n' = \int_0^R r^n F(r)\,dr,$$

are known as the moments about the origin of the frequency functions $\phi(x)$ and $F(r)$. As is well known from the papers of Professor Karl Pearson, it is generally possible to construct the frequency function, when the numerical values of a sufficient number of the moments have been obtained. As a rule the first four moments, in exceptional cases the first six moments, are sufficient.

Now, the moments v_n' of $\phi(x)$ may be numerically determined from the observed distribution of the apparent diameters. The mechanical quadrature formulae for this determination need not be entered into here, as they are familiar to any statistician. Hence the solution of (5) is reduced to the following problem: To express the moments N_n' of the sought function $F(r)$ in terms of the moments v_n'.

---

* von Zeipel has used this formula in his cluster reductions, where the accuracy required is much greater than in the case of the anatomical applications.

Indeed, one such relation may be written out at once. As the plane section contains Nr_0 corpuscle images per unit area (see § 3), it is clear that $N . N_1' \frac{\pi}{4} v_2'$ (the mean diameter r_0 being here denoted by N_1') is the area of the images per unit area in the sections. Furthermore $N \frac{\pi}{6} N_3'$ is the volume of the corpuscles per unit volume of the tissue. As these two quantities are evidently to be put equal to each other, we have

$$N_3' = \tfrac{3}{2} N_1' v_2'.$$

A general solution is obtained as follows: We have, putting $r_0 = N_1'$ in (5),

$$\phi(x) = \frac{x}{N_1'} \int_x^R F(r) \frac{dr}{\sqrt{r^2 - x^2}}.$$

Hence
$$v_n' = \frac{1}{N_1'} \int_0^R x^{n+1}\, dx \int_x^R F(r) \frac{dr}{\sqrt{r^2 - x^2}}.$$

Using Dirichlet's formula* this reduces to

$$v_n' = \frac{1}{N_1'} \int_0^R F(r)\, dr \int_0^r \frac{x^{n+1}}{\sqrt{r^2 - x^2}} dx.$$

Putting
$$y_n = \int_0^r \frac{x^{n+1}}{\sqrt{r^2 - x^2}} dx,$$

it follows from well-known theorems in the theory of integration, that we have

$$y_0 = r; \quad y_1 = \frac{\pi}{4} r^2; \quad y_2 = \tfrac{2}{3} r^3; \quad y_3 = \frac{3\pi}{16} r^4; \quad y_4 = \tfrac{8}{15} r^5; \quad y_5 = \frac{5\pi}{32} r^6; \quad y_6 = \tfrac{16}{35} r^7.$$

Hence we finally get

$$\left.\begin{array}{l} N_1' = N_1' v_0' \text{ (i.e., as should be, } v_0' = 1), \\ N_2' = \dfrac{4}{\pi} N_1' v_1'\,; \quad N_3' = \tfrac{3}{2} N_1' v_2'\,; \quad N_4' = \dfrac{16}{3\pi} N_1' v_3'\,; \\ N_5' = \tfrac{15}{8} N_1' v_4'\,; \quad N_6' = \dfrac{32}{5\pi} N_1' v_5'\,; \quad N_7' = \tfrac{35}{16} N_1' v_6'. \end{array}\right\} \quad \ldots\ldots\ldots (11)$$

It thus only remains to determine N_1', and the problem to express the "true moments" in terms of the "apparent moments" is solved.

In order to determine N_1' we may for instance make use of the proposition: *The arithmetic mean of the true diameters is equal to the harmonic mean of the apparent diameters multiplied by* $\pi/2$. The proof of this proposition is evidently

$$N_1 \int_0^R \phi(x) \frac{dx}{x} = \int_0^R dx \int_x^R F(r) \frac{dr}{\sqrt{r^2 - x^2}} = \int_0^R F(r)\, dr \int_0^r \frac{dx}{\sqrt{r^2 - x^2}} = \frac{\pi}{2} \int_0^R F(r)\, dr = \frac{\pi}{2}.$$

The moments of the sought function $F(r)$ thus being obtained, the construction of the function only depends on a suitable choice of an analytical expression. I believe that Pearson's theory of frequency curves and the auxiliary tables given by him for the construction work, give ample means to meet any requirements in this respect.

---

* **As to the applicability of Dirichlet's formula when the integrand becomes infinite at the limits of integration see Bôcher, *loc. cit.* p. 4.**

It may be added, that the general relations between the moments N' and v' may be expressed as follows:

$$V_1'v_n' = \begin{cases} N'_{n+1}\dfrac{1.3.5\ldots n}{2.4.6\ldots(n+1)}\cdot\dfrac{\pi}{2}, & \text{when } n \text{ is odd,} \\ N'_{n+1}\dfrac{2.4.6\ldots n}{3.5.7\ldots(n+1)}, & \text{when } n \text{ is even.} \end{cases}$$

6. In § 2 it was pointed out that under certain conditions the true distribution $F(r)$ and the apparent distribution $\phi(x)$ become identical. In order to show this we put in (5) $\phi(x) \equiv F(x)$, thus getting the equation of condition

$$F(x) = \frac{x}{r_0}\int_x^R F(r)\frac{dr}{\sqrt{r^2-x^2}}.$$

Putting $F(x) : x = F_1(x)$, we get instead

$$F_1(x) = \frac{1}{r_0}\int_x^R F_1(r)\frac{r\,dr}{\sqrt{r^2-x^2}} = \frac{1}{r_0}\int_0^{\sqrt{R^2-x^2}} F_1(\sqrt{x^2+y^2})\,dy.$$

It is seen at once that this equation is satisfied by the function

$$F_1(x) = ke^{-\frac{x^2}{2\sigma^2}},$$

if, as in this case must be done, we put $R = \infty$. Hence as we must also have

$$\int_0^\infty F(r)\,dr = 1,$$

we get
$$F(r) = \frac{1}{\sigma^2}re^{-\frac{r^2}{2\sigma^2}},$$

and
$$r_0 = \sigma\sqrt{\frac{\pi}{2}}.$$

Our conclusion is that if the true distribution has the form

$$F(r) = \frac{r}{\sigma^2}e^{-\frac{r^2}{2\sigma^2}},$$

then the apparent distribution has the same form, or

$$\phi(x) = \frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}.$$

That this result is true also in the inverse sense may be shown by the aid of the solution (8) of (5).

It might at first seem as if the same property is at hand when we have the more general form

$$\phi(x) = \sum_1^n k_n\frac{x}{\sigma_n^2}e^{-\frac{x^2}{2\sigma_n^2}}.$$

It is easily found, however, that in this case we get

$$F(r) = \sum_1^n k_n \frac{r}{\sigma_n^3} e^{-\frac{r^2}{2\sigma_n^2}} : \sum_1^n \frac{k_n}{\sigma_n}.$$

7. In § 3 we denoted by

$$f(r) = \frac{rF(r)}{r_0}$$

the relative frequency function of the true diameters of the corpuscles, represented in a random section. Introducing this function the equation (5) takes the form

$$\phi(x) = x \int_x^R f(r) \frac{dr}{r\sqrt{r^2 - x^2}}.$$

In this form the equation is valid also if the corpuscles are not spherical but of ellipsoidal shape. The different corpuscles need not be of the same form, nor must the axes of the different ellipsoids be parallel, provided that as "diameter" of a plane section out of any corpuscle we take the *geometric mean* of its largest and smallest diameters. The apparent diameter x of a corpuscle should thus be defined as the geometric mean of the largest and smallest diameters in the image obtained in a random section, the true diameter r as the geometric mean of the largest and smallest diameters in the central plane parallel to the section plane.

This rather remarkable fact follows from the following proposition: Take any three-axial ellipsoid and place it with the principal axes pointing in any direction. Two horizontal sections are taken in this ellipsoid, one passing through the centre, the other passing at the distance y from the centre. The areas of the two sections we denote by A_0 and A_y. The distance of the horizontal tangent plane from the centre we denote by h. Then we always have

$$A_y = A_0\left(1 - \frac{y^2}{h^2}\right).$$

The truth of this proposition follows immediately from the property of an ellipsoid, that the areas of parallel sections are proportional to the areas of similarly situated sections in a sphere.

Now, defining the apparent and the true diameters in the above way, we have

$$\pi r^2 = 4A_0; \quad \pi x^2 = 4A_y,$$

and it is seen that we have the relation

$$x = r\sqrt{1 - \frac{y^2}{h^2}},$$

or

$$\frac{y}{h} = \frac{\sqrt{r^2 - x^2}}{r}$$

But

$$\frac{dy}{h\,dx}\,dx = \frac{x\,dx}{r\sqrt{r^2 - x^2}}$$

is the probability that an arbitrary section in a corpuscle of central "diameter" r has an apparent "diameter" x to $x + dx$. Hence this probability is identically the same as in the case of a sphere.

8. It was mentioned above, that the "corpuscle problem" has a parallel in stellar statistics. This parallel is the "globular cluster problem."

The Globular Cluster Problem. A globular cluster, i.e. an agglomeration of stars in spherical layers of equal density around a common centre, is projected on the heavens (or a photographic plate). To express the "true" law of density in space in terms of the observed "apparent" law of density in the projection.

Let $f(r)\,dr$ be the relative number of stars at the distance r to $r+dr$ from the centre of the cluster, and let $\phi(x)\,dx$ be the relative number of stars projected on a plane and falling within the circular ring of radius x to $x+dx$. If $\Delta(r)$ is the density function of the cluster and $\delta(x)$ is the density function of the projection we evidently have

$$N.f(r) = 4\pi r^2\,\Delta(r)\,;\quad N\phi(x) = 2\pi x\delta(x),$$

N being the total number of stars in the cluster. But it is immediately seen that we must have (R being the superior limit to the radius vector)

$$\delta(x) = 2\int_0^{\sqrt{R^2-x^2}} \Delta(\sqrt{x^2+y^2})\,dy,$$

or by a simple transformation

$$\delta(x) = 2\int_x^R \Delta(r)\,\frac{r\,dr}{\sqrt{r^2-x^2}}.$$

Hence we get

$$\phi(x) = x\int_x^R f(r)\,\frac{dr}{r\sqrt{r^2-x^2}},$$

which is the same functional relation as in the corpuscle problem. The two problems are thus—although different in origin and formulation—mathematically identical.

9. The methods of solution expanded in §§ 4 and 5 involve computations and constructions, which will certainly be easily mastered by a trained statistician. I am afraid, however, that the anatomists, who generally have neither the mathematical nor the arithmetical training necessary, will find the methods too complicated for application. Furthermore I suppose that the computations involved will be too unwieldy, when it is required to reduce the observations from numerous organs (the material of Hellman embraces several hundred spleens). To obtain a more ready method of reduction I have proceeded as follows:

The observations give directly the distribution of the apparent diameters in the classes 1·5—2·5; 2·5—3·5; 3·5—4·5 mm., etc. We will denote the frequencies in these classes by $f_2, f_3, f_4, f_5, \ldots f_R$. Although diameters in the classes 0—0·5 and 0·5—1·5 are generally not observed and hence the frequencies in these classes will be unknown, we may, for the sake of completeness, take account of them in the following developments, denoting them by $f_{\frac{1}{2}}$ and f_1. We now have, of course,

$$f_{\frac{1}{2}} = kr_0\int_0^{\frac{1}{2}} \phi(x)\,dx\,;\quad f_i = kr_0\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \phi(x)\,dx \quad \ldots\ldots\ldots\ldots\ldots\ldots(12).$$

The constant k, which as a factor of proportionality will disappear in the final reductions to relative frequencies, here assumes the value NY, where Y is the total area of the observed sections, and N, as before, the number of corpuscles to the unit volume of tissue.

From equation (5) we now have ($i = 1, 2, \ldots$)

$$f_i = k\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} x\,dx \int_x^R F(r)\,\frac{dr}{\sqrt{r^2-x^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(13),$$

or, changing the order of integration and integrating over the variable x,

$$f_i = k\int_{i-\frac{1}{2}}^{i+\frac{1}{2}} dr\,F(r)\sqrt{r^2-(i-0{\cdot}5)^2}$$
$$+\,k\int_{i+\frac{1}{2}}^{R} dr\,F(r)\left(\sqrt{r^2-(i-0{\cdot}5)^2}-\sqrt{r^2-(i+0{\cdot}5)^2}\right) \ldots(14).$$

For $f_{\frac{1}{2}}$ we get

$$f_{\frac{1}{2}} = k\int_0^{\frac{1}{2}} dr\,F(r).\,r + k\int_{\frac{1}{2}}^{R} dr\,F(r)\left(r-\sqrt{r^2-0{\cdot}25}\right) \ldots\ldots\ldots\ldots(15).$$

As only sporadic corpuscles are found with a diameter above 15 mm. we may put $R = 15{\cdot}5$. Furthermore, we put

$$P_{\frac{1}{2}} = k\int_0^{\frac{1}{2}} dr\,F(r); \quad P_n = \int_{n-\frac{1}{2}}^{n+\frac{1}{2}} dr\,F(r),$$

and it will be seen from the systems (14, 15), that if we put

$$f_{\frac{1}{2}} = P_{\frac{1}{2}}a_{00} + \sum_{j=1}^{15} P_j a_{0j}, \quad f_i = \sum_{j=0}^{15-i} P_{i+j}a_{ij} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(16),$$

the coefficients a_{ij} will be approximately given by

$$a_{00} = 0{\cdot}25; \quad a_{0j} = j-\sqrt{j^2-0{\cdot}25}; \quad a_{i0} = \sqrt{i^2-(i-0{\cdot}5)^2};$$
$$a_{ij} = \sqrt{(i+j)^2-(i-0{\cdot}5)^2}-\sqrt{(i+j)^2-(i+0{\cdot}5)^2} \ldots\ldots\ldots\ldots(16^{bis}).$$

Now, the system (16) is an approximative system of equations. The nature of the approximations involved will be clear if it is observed that the system has the following "exact" meaning: If determined from the system (16) with the coefficients (16^{bis}) the quantities $P_{\frac{1}{2}}$; P_1; P_2; ... P_{15} are proportional to the numbers of corpuscles having their diameters exactly equal to ½; 1; 2; 3; ... 15 mm. respectively, which, if cut by a random plane section, will give the observed distribution $f_{\frac{1}{2}}$; f_1; f_2; ... f_{15} of apparent diameters. Geometrical considerations indeed show that the coefficients a_{ij} are numerically equal to the combined thickness of the two layers parallel to the section plane, in which the centres of corpuscles of diameter j must lie, in order that the apparent diameter be between $i - 0{\cdot}5$ and $i + 0{\cdot}5$.

The use of the system (18) will of course be the following: Inserting the observed values of f_i the corresponding values of P_i are obtained. When, as in the material of Hellman is the case, we do not know $f_{\frac{1}{2}}$ and f_1, it is seen that we can still use all the equations (18) except the first two. Having obtained the values of

Determining the numerical values of the coefficients, the system (16) becomes:

$$\begin{array}{r}
0{\cdot}2500P_{\frac{1}{2}}+0{\cdot}1340P_1+0{\cdot}0635P_2+0{\cdot}0420P_3+0{\cdot}0314P_4+0{\cdot}0251P_5+0{\cdot}0209P_6+0{\cdot}0179P_7+0{\cdot}0156P_8+0{\cdot}0139P_9+0{\cdot}0125P_{10}+0{\cdot}0114P_{11}+0{\cdot}0104P_{12}+0{\cdot}0096P_{13}+0{\cdot}0089P_{14}+0{\cdot}0084P_{15}=f_{\frac{1}{2}}\\
0{\cdot}8660P_1+0{\cdot}6136P_2+0{\cdot}3599P_3+0{\cdot}2605P_4+0{\cdot}2052P_5+0{\cdot}1696P_6+0{\cdot}1447P_7+0{\cdot}1263P_8+0{\cdot}1120P_9+0{\cdot}1006P_{10}+0{\cdot}0914P_{11}+0{\cdot}0838P_{12}+0{\cdot}0772P_{13}+0{\cdot}0717P_{14}+0{\cdot}0668P_{15}=f_1\\
1{\cdot}3229P_2+0{\cdot}9398P_3+0{\cdot}5856P_4+0{\cdot}4396P_5+0{\cdot}3551P_6+0{\cdot}2991P_7+0{\cdot}2588P_8+0{\cdot}2283P_9+0{\cdot}2044P_{10}+0{\cdot}1851P_{11}+0{\cdot}1691P_{12}+0{\cdot}1558P_{13}+0{\cdot}1444P_{14}+0{\cdot}1346P_{15}=f_2\\
1{\cdot}6583P_3+1{\cdot}1860P_4+0{\cdot}7594P_5+0{\cdot}5810P_6+0{\cdot}4761P_7+0{\cdot}4056P_8+0{\cdot}3542P_9+0{\cdot}3150P_{10}+0{\cdot}2838P_{11}+0{\cdot}2585P_{12}+0{\cdot}2374P_{13}+0{\cdot}2197P_{14}+0{\cdot}2042P_{15}=f_3\\
1{\cdot}9365P_4+1{\cdot}3913P_5+0{\cdot}9048P_6+0{\cdot}7003P_7+0{\cdot}5793P_8+0{\cdot}4974P_9+0{\cdot}4372P_{10}+0{\cdot}3909P_{11}+0{\cdot}3539P_{12}+0{\cdot}3237P_{13}+0{\cdot}2982P_{14}+0{\cdot}2769P_{15}=f_4\\
2{\cdot}1794P_5+1{\cdot}5707P_6+1{\cdot}0318P_7+0{\cdot}8049P_8+0{\cdot}6703P_9+0{\cdot}5787P_{10}+0{\cdot}5111P_{11}+0{\cdot}4589P_{12}+0{\cdot}4171P_{13}+0{\cdot}3827P_{14}+0{\cdot}3538P_{15}=f_5\\
2{\cdot}3979P_6+1{\cdot}7320P_7+1{\cdot}1458P_8+0{\cdot}8990P_9+0{\cdot}7523P_{10}+0{\cdot}6522P_{11}+0{\cdot}5783P_{12}+0{\cdot}5209P_{13}+0{\cdot}4748P_{14}+0{\cdot}4368P_{15}=f_6\\
2{\cdot}5981P_7+1{\cdot}8798P_8+1{\cdot}2500P_9+0{\cdot}9849P_{10}+0{\cdot}8274P_{11}+0{\cdot}7196P_{12}+0{\cdot}6399P_{13}+0{\cdot}5780P_{14}+0{\cdot}5281P_{15}=f_7\\
2{\cdot}7839P_8+2{\cdot}0169P_9+1{\cdot}3466P_{10}+1{\cdot}0646P_{11}+0{\cdot}8970P_{12}+0{\cdot}7822P_{13}+0{\cdot}6973P_{14}+0{\cdot}6312P_{15}=f_8\\
2{\cdot}9580P_9+2{\cdot}1453P_{10}+1{\cdot}4368P_{11}+1{\cdot}1391P_{12}+0{\cdot}9621P_{13}+0{\cdot}8408P_{14}+0{\cdot}7510P_{15}=f_9\\
3{\cdot}1225P_{10}+2{\cdot}2666P_{11}+1{\cdot}5219P_{12}+1{\cdot}2092P_{13}+1{\cdot}0234P_{14}+0{\cdot}8961P_{15}=f_{10}\\
3{\cdot}2787P_{11}+2{\cdot}3817P_{12}+1{\cdot}6027P_{13}+1{\cdot}2757P_{14}+1{\cdot}0814P_{15}=f_{11}\\
3{\cdot}4278P_{12}+2{\cdot}4915P_{13}+1{\cdot}6796P_{14}+1{\cdot}3391P_{15}=f_{12}\\
3{\cdot}5707P_{13}+2{\cdot}5967P_{14}+1{\cdot}7533P_{15}=f_{13}\\
3{\cdot}7081P_{14}+2{\cdot}6977P_{15}=f_{14}\\
3{\cdot}8406P_{15}=f_{15}\\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(17).
\end{array}$$

As the quantities P are here the unknowns and the quantities f may be regarded as known, we must solve this system of linear equations. The solution is found to be:

$$\begin{array}{r}
4{\cdot}0000f_{\frac{1}{2}}-0{\cdot}6189f_1+0{\cdot}0951f_2-0{\cdot}0209f_3+0{\cdot}0024f_4-0{\cdot}0012f_5-0{\cdot}0002f_6-0{\cdot}0002f_7-0{\cdot}0001f_8-0{\cdot}0001f_9-0{\cdot}0001f_{10}-0{\cdot}0001f_{11}-0{\cdot}0000f_{12}-0{\cdot}0000f_{13}-0{\cdot}0000f_{14}-0{\cdot}0000f_{15}=P_{\frac{1}{2}}\\
1{\cdot}1547f_1-0{\cdot}5356f_2+0{\cdot}0529f_3-0{\cdot}0258f_4-0{\cdot}0027f_5-0{\cdot}0037f_6-0{\cdot}0019f_7-0{\cdot}0014f_8-0{\cdot}0009f_9-0{\cdot}0007f_{10}-0{\cdot}0005f_{11}-0{\cdot}0004f_{12}-0{\cdot}0003f_{13}-0{\cdot}0003f_{14}-0{\cdot}0002f_{15}=P_1\\
0{\cdot}7559f_2-0{\cdot}4284f_3+0{\cdot}0338f_4-0{\cdot}0248f_5-0{\cdot}0047f_6-0{\cdot}0047f_7-0{\cdot}0026f_8-0{\cdot}0019f_9-0{\cdot}0014f_{10}-0{\cdot}0010f_{11}-0{\cdot}0007f_{12}-0{\cdot}0006f_{13}-0{\cdot}0005f_{14}-0{\cdot}0004f_{15}=P_2\\
0{\cdot}6030f_3-0{\cdot}3693f_4+0{\cdot}0256f_5-0{\cdot}0236f_6-0{\cdot}0054f_7-0{\cdot}0051f_8-0{\cdot}0030f_9-0{\cdot}0022f_{10}-0{\cdot}0016f_{11}-0{\cdot}0012f_{12}-0{\cdot}0009f_{13}-0{\cdot}0008f_{14}-0{\cdot}0006f_{15}=P_3\\
0{\cdot}5164f_4-0{\cdot}3297f_5+0{\cdot}0211f_6-0{\cdot}0223f_7-0{\cdot}0057f_8-0{\cdot}0052f_9-0{\cdot}0032f_{10}-0{\cdot}0024f_{11}-0{\cdot}0017f_{12}-0{\cdot}0014f_{13}-0{\cdot}0010f_{14}-0{\cdot}0009f_{15}=P_4\\
0{\cdot}4588f_5-0{\cdot}3006f_6+0{\cdot}0181f_7-0{\cdot}0212f_8-0{\cdot}0058f_9-0{\cdot}0052f_{10}-0{\cdot}0033f_{11}-0{\cdot}0025f_{12}-0{\cdot}0018f_{13}-0{\cdot}0014f_{14}-0{\cdot}0011f_{15}=P_5\\
0{\cdot}4170f_6-0{\cdot}2780f_7+0{\cdot}0161f_8-0{\cdot}0202f_9-0{\cdot}0058f_{10}-0{\cdot}0051f_{11}-0{\cdot}0033f_{12}-0{\cdot}0025f_{13}-0{\cdot}0019f_{14}-0{\cdot}0015f_{15}=P_6\\
0{\cdot}3849f_7-0{\cdot}2600f_8+0{\cdot}0146f_9-0{\cdot}0193f_{10}-0{\cdot}0058f_{11}-0{\cdot}0050f_{12}-0{\cdot}0033f_{13}-0{\cdot}0025f_{14}-0{\cdot}0019f_{15}=P_7\\
0{\cdot}3592f_8-0{\cdot}2449f_9+0{\cdot}0134f_{10}-0{\cdot}0185f_{11}-0{\cdot}0057f_{12}-0{\cdot}0050f_{13}-0{\cdot}0033f_{14}-0{\cdot}0025f_{15}=P_8\\
0{\cdot}3381f_9-0{\cdot}2323f_{10}+0{\cdot}0124f_{11}-0{\cdot}0179f_{12}-0{\cdot}0056f_{13}-0{\cdot}0049f_{14}-0{\cdot}0032f_{15}=P_9\\
0{\cdot}3203f_{10}-0{\cdot}2214f_{11}+0{\cdot}0116f_{12}-0{\cdot}0172f_{13}-0{\cdot}0054f_{14}-0{\cdot}0048f_{15}=P_{10}\\
0{\cdot}3050f_{11}-0{\cdot}2119f_{12}+0{\cdot}0110f_{13}-0{\cdot}0166f_{14}-0{\cdot}0053f_{15}=P_{11}\\
0{\cdot}2917f_{12}-0{\cdot}2036f_{13}+0{\cdot}0104f_{14}-0{\cdot}0161f_{15}=P_{12}\\
0{\cdot}2801f_{13}-0{\cdot}1961f_{14}+0{\cdot}0099f_{15}=P_{13}\\
0{\cdot}2697f_{14}-0{\cdot}1894f_{15}=P_{14}\\
0{\cdot}2604f_{15}=P_{15}\\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(18).
\end{array}$$

$P_2, P_3, \ldots P_{15}$ it will, however, be possible to find the value of $P_{\frac{1}{2}}$ and P_1 by a graphical extrapolation—or rather interpolation, as the curve may be assumed to pass through the origin. Finally, dividing the quantities P by their sum S we have

$$F_{\frac{1}{2}} = \frac{P_{\frac{1}{2}}}{S}, \quad F_i = \frac{P_i}{S},$$

where $F_{\frac{1}{2}}, F_1, F_2, F_3, \ldots F_{15}$ denote the relative frequencies of the true diameters in the classes 0—0·5, 0·5—1·5, 1·5—2·5, 2·5—3·5, ... 14·5—15·5.

In the next section we shall give some numerical examples of the use of equations (18).

10. *Numerical examples.* Of the following numerical examples the first three are arranged in order to obtain a practical test of the mechanical solution (18). In these examples the exact answer is known beforehand, and the examples have been so chosen, that they illustrate the applicability of the method in some typical cases.

Example 1. As a first example I take the following distribution:

$$\phi(x) = \frac{x}{(5)^2} e^{-\frac{x^2}{2(5)^2}}.$$

We here evidently have in *pro mille*

$$f_{\frac{1}{2}} = 1000 \int_0^{\frac{1}{2}} \frac{x}{25} e^{-\frac{x^2}{50}} dx = 1000 \left(1 - e^{-\frac{(0\cdot5)^2}{50}}\right),$$

and

$$f_i = 1000 \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \frac{x}{25} e^{-\frac{x^2}{50}} dx = 1000 \left(e^{-\frac{(i-\frac{1}{2})^2}{50}} - e^{-\frac{(i+\frac{1}{2})^2}{50}}\right) \quad \ldots\ldots\ldots(19).$$

As was shown in § 6 the true distribution is here

$$F(r) = \frac{r}{5^2} e^{-\frac{r^2}{2\cdot5^2}};$$

and consequently, if the system (18) is applied, the $F_i (= P_i : \Sigma P_i)$ found should be approximately equal to the corresponding f_i.

From (19) we get, writing the numerical values below the symbols,

$f_{\frac{1}{2}}$	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}		Σf
5	39	73	100	116	121	117	105	89	71	54	39	27	18	11	7	8	1000

The residual number (= 8) is of course the value of the integral

$$\int_{15\cdot5}^{\infty} \phi(x)\, dx.$$

It may here be left entirely out of consideration.

Inserting the above values of the f_i in the system (18) we get

$P_{\frac{1}{2}}$	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}
0·8	6·9	11·9	16·3	19·0	19·7	19·0	16·9	14·3	11·3	8·6	6·2	4·2	2·9	1·7	1·8

of which the sum is 161·5. Expressing this in *pro mille* we get

$F_{\frac{1}{2}}$	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}	F_{15}	ΣF
5	43	74	101	118	122	118	105	89	70	53	38	26	18	11	11	1000+2

It is seen that these values of F_i are, indeed, very nearly equal to the f_i, which shows that in the case treated the system (18) has worked very well. That there should be slight deviations at the extremes was, of course, to be expected.

Finally, it follows from § 6 that here we have

$$r_0 = 5 \,.\, \sqrt{\frac{\pi}{2}} = 6\cdot26.$$

It follows, however, from § 9 that we also have

$$\frac{1}{r_0} = \frac{\Sigma P_i}{\Sigma f_i} = 0\cdot1608,$$

which gives in this case

$$r_0 = 6\cdot22.$$

Example 2. It is evident that the system (18) can be used also when we have fewer that 15 classes. We then must of course put the superfluous f_i equal to zero. It is also clear, however, that the fewer the classes the less exact will the mechanical solution be. In order to test the degree of approximation in a rather extreme case of 6 classes only, we shall apply the method to the distribution

$$\phi(x) = \frac{x}{(1\cdot27)^2}\, e^{-\frac{x^2}{2(1\cdot27)^2}},$$

which gives in *pro mille*

$f_{\frac{1}{2}}$	f_1	f_2	f_3	f_4	f_5	Σf
73	428	355	121	20	2	1000 − 1.

Equations (18) now give

$P_{\frac{1}{2}}$	P_1	P_2	P_3	P_4	P_5	ΣP
57	309	218	64	10	1·0	659

or in *pro mille*

$F_{\frac{1}{2}}$	F_1	F_2	F_3	F_4	F_5	ΣF
87	468	331	98	15	1·5	1000·5.

As an exact solution would, here also, give the apparent distribution back again, we find that 6 classes are generally too few if a higher degree of approximation is required. For the anatomic praxis here in view the accuracy required is, however, not so great but that the application of (18) to series of 6 classes will give fairly satisfactory results.

The exact value of r_0 is here

$$r_0 = 1\cdot27 \sqrt{\frac{\pi}{2}} = 1\cdot57.$$

The approximate solution gives

$$r_0 = \tfrac{1000}{659} = 1\cdot52.$$

Example 3. The accuracy of the solution (18) depends not only on the number of classes into which the observations have been divided, but will be affected also by extreme skewness or abnormal excess (i.e. abnormal clustering in the central classes) in the observed distribution. As an example of a distribution of this kind we may take the following one, derived from the frequency function

$$\phi(x) = 0{\cdot}65\, \frac{x}{(1{\cdot}27)^2}\, e^{-\frac{x^2}{2(1{\cdot}27)^2}} + 0{\cdot}35\, \frac{x}{(4{\cdot}35)^2}\, e^{-\frac{x^2}{2(4{\cdot}35)^2}}.$$

This function has actually been found by me to express the apparent distribution of the stars within the stellar cluster *Messier* 3.

The function gives

$f_{\frac{1}{2}}$	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	$f\Sigma$
50	295	264	124	59	49	43	36	28	20	13	8	5	3	1000 − 3

and we find from (18)

$P_{\frac{1}{2}}$	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	ΣP
39·9	204·1	146·8	52·9	14·1	9·5	7·9	6·7	5·3	3·8	2·4	1·3	0·9	0·8	496·4

or in *pro mille*

$F_{\frac{1}{2}}$	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	ΣF
80	409	296	107	28	19	16	14	11	8	5	3	2	2	1000

Thus $$r_0 = \frac{1000}{496{\cdot}4} = 2{\cdot}01.$$

The exact distribution is here (§ 6)

$$F_i = 1692 \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} \left(0{\cdot}65\, \frac{r}{(1{\cdot}27)^3}\, e^{-\frac{r^2}{2(1{\cdot}27)^2}} + 0{\cdot}35\, \frac{r}{(4{\cdot}35)^3}\, e^{-\frac{r^2}{2(4{\cdot}35)^2}}\right) dr,$$

which gives

$$r_0 = 1{\cdot}692 \sqrt{\frac{\pi}{2}} = 2{\cdot}12,$$

and

$F_{\frac{1}{2}}$	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}
64	377	320	123	35	20	16	14	10	7	5	3	2	1

It is seen that the result of equations (18) is here somewhat inexact for the classes where the apparent diameters are most clustered. This approximation would, however, be quite sufficient for anatomical purposes.

Example 4. As a fourth example we take the following distribution, observed by Hellman in 8 sections from a human spleen. (The diameters are given in mm., as measured with 18 × 18 magnification.)

2·5—3·5	3·5—4·5	4·5—5·5	5·5—6·5	6·5—7·5	7·5—8·5	8·5—9·5	9·5—10·5
$f_3 = 146$	$f_4 = 197$	$f_5 = 210$	$f_6 = 184$	$f_7 = 143$	$f_8 = 95$	$f_9 = 57$	$f_{10} = 31$

10·5—11·5	11·5—12·5	12·5—13·5	13·5—14·5	14·5—1·55	Sum
$f_{11} = 15$	$f_{12} = 7$	$f_{13} = 4$	$f_{14} = 2$	$f_{15} = 1$	1092

The frequencies have been slightly smoothed graphically. For the classes 0—0·5 and 0·5—1·5 no observations have been made; for the class 1·5—2·5 the frequency $f_2 = 52$ was found, but this value is evidently too small (it would give a negative value of P_2), probably owing to the fact that the observations are not complete down to 1·5 mm.

Using, consequently, only f_3 to f_{15} we find, inserting in (18),

P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	ΣP
15·4	32·1	41·5	36·7	30·4	20·4	12·1	6·6	3·1	1·3	0·6	0·4	0·3	200·9

Plotting these values graphically, we may venture an interpolation of the values of $P_{\frac{1}{2}}$, P_1 and P_2, in continuing the curve down to the origin. We have thus found $P_{\frac{1}{2}} = 1$, $P_1 = 3$, and $P_2 = 8$. The sum now becomes 212·9 and we get *pro mille*

$F_{\frac{1}{2}}$	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}	F_{15}
5	14	38	72	151	194	172	143	96	57	31	15	6	3	2	1

This gives

$$r_0 = \frac{1}{1000}\left(\sum_{1}^{15} r F_r + 0{\cdot}25 F_{\frac{1}{2}}\right) = 5{\cdot}82$$

for the mean value of the true diameter.

THE FURTHER THEORY OF FRANCIS GALTON'S INDIVIDUAL DIFFERENCE PROBLEM.

By J. O. IRWIN, M.A., M.Sc.

CONTENTS.

	PAGE
Introduction	100
Mean Difference between the pth and qth Individuals in Order of Magnitude of a given Character	102
Moments of the Frequency Distribution of Differences between Consecutive Individuals	104
Actual Frequency Distribution of Differences between consecutive Individuals	108
Approximations to the Frequency Distribution of Differences between Consecutive Individuals	116
Conclusion, including Consideration of the Probable Errors of Correlation Coefficients calculated by use of the Method	120

INTRODUCTION.

In August 1902 Sir Francis Galton published a memoir entitled "The Most Suitable Proportion between the Values of First and Second Prizes." He concluded that if the amount of the prizes had to be decided before the competition, it seemed reasonable to divide the prize money in the ratio of the mean excess in merit (supposing it measurable) of candidates who are first over those who are second to the mean excess of those who are second over those who are third.

This led Professor Pearson to consider the following problem:

"A random sample of n individuals is taken from a population of N members, which when N is very large may be taken to obey some law of frequency which is expressed by the curve $y = N\phi(x)$, $y\,\delta x$ being the total frequency of individuals with characters or organs lying between x and $x + \delta x$. It is required to find an expression for the average difference in character between the pth and $(p+1)$th individuals when the sample is arranged in order of magnitude of the character."

This general problem was termed by him "Francis Galton's Individual Difference Problem in Statistics." If $\chi_{n,p}$ be the average difference it is required to find, the solution of the problem is given by

$$\chi_{n,p} = \frac{n!}{(n-p)!\,p!}\int_{-\infty}^{\infty} \alpha^{n-p}(1-\alpha)^p\,dx,$$

where

$$\alpha = \int_{-\infty}^{x} \phi(x)\,dx \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}.$$

This result was obtained independently by Professor Pearson and Dr Sheppard and some numerical results are given in the original paper*, on the assumption that the original population is normally distributed when (i) becomes

$$\chi_{n,p} = \frac{\sigma . n!}{(n-p)!\, p!} \int_{-\infty}^{\infty} \alpha^{n-p} (1-\alpha)^p \, dx,$$

where $$\alpha = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \quad \text{..............................(ii)},$$

σ being the standard deviation of the original population, and particular values of α being obtainable from the tables of the probability integral.

Two methods were given for the computation:

(α) Quadrature.

(β) An expansion of the integral in a series about the mode of the integrand by a method due to Laplace.

The problem was regarded as a new and important departure in statistical theory, in fact Professor Pearson concluded his investigation in these words: "It (the method) opens up, indeed, many new methods of enquiry, the effectiveness of which however can only be tested by their application in actual statistical practice†."

In order to consider the likelihood of the method being useful in statistics, it is necessary to go rather further into the theory than has hitherto been attempted.

It seems desirable to find not only the mean difference between the pth and $(p+1)$th individuals, but also the standard deviation of that difference, and even the actual frequency distribution of differences, since if it should turn out to be

---

* *Biometrika*, Vol. I. pp. 385–399.

† It was suggested by Professor Pearson that the above result might be used to obtain fraternal correlation coefficients in the following manner:

Pairs of individuals are drawn at random from the general population and their mean difference for the character in question is measured. Hence from (i) we may find

$$\chi_{2,1} = \lambda\sigma,$$

σ being the standard deviation of the original population, λ a known number calculated from (i) and $\chi_{2,1}$ the mean of the differences between the pairs drawn. If now pairs of individuals be taken not quite at random, but from correlated groups, e.g. pairs of brothers taken at random, the result will be λ times the standard deviation of the correlated groups, e.g. a group of brethren, or we have

$$\chi'_{2,1} = \lambda\sigma\sqrt{1-r^2},$$

where r is the correlation of the group, e.g. the correlation between brothers.

Whence $$\sqrt{1-r^2} = \frac{\chi'_{2,1}}{\chi_{2,1}}.$$

Now if the phrase "pairs of brothers taken at random" means "brothers" in the ordinary sense, e.g. pairs of individuals who are actually brothers of one another, I do not think this expression for r is correct. It would however be correct on the assumption of linear regression and homoscedasticity, if by the term "brothers" we meant individuals each of whom had a brother with the same value of the character, i.e. if we took pairs of individuals at random, each pair from the same array of the correlation table.

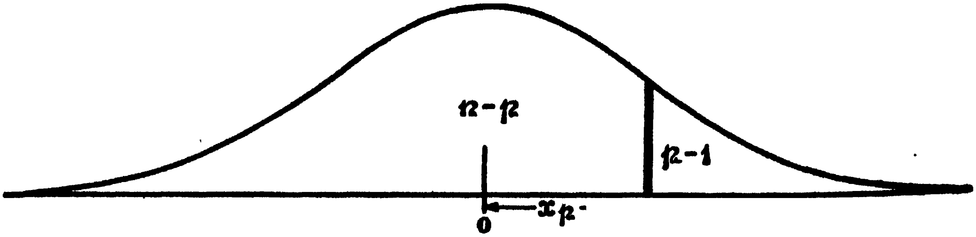
But this is of no practical importance as if we had enough material to form a correlation table we should use the product-moment method, certainly no one would under these circumstances average the differences between pairs taken from the same array.

At the end of this paper, p. 126, we give the correct expression for r when we average differences between pairs of "brothers," in the ordinary sense of the word.

markedly different from the Gaussian that fact would have to be taken into account in any applications of the method. It is the object of this paper to determine that frequency distribution approximately and if possible exactly, on the hypothesis that the original population is normally distributed.

(1) *Mean Difference between the* p*th and* q*th Individuals in Order of Magnitude of a given Character.*

Although not directly concerned with the main point of this investigation, it may be of interest to have a direct formula for the mean difference between the pth and qth individuals in random samples of n, arranged in order of magnitude of the character.



Assuming a frequency distribution $y = N\phi(x)$ in our original population of N (where N is very large), let x_p, x_q be the magnitude of the characters of the pth and qth individuals; then

$$\alpha = \int_{-\infty}^{x_p} \phi(x)\,dx = \text{chance of an individual having a value of the character} < x_p,$$

and $1 - \alpha =$ chance of an individual with a greater value of the character.

Therefore chance of the pth individual having his value of the character between x_p and $x_p + dx_p$

$$= \frac{n!}{(n-p)!\,(p-1)!}\,\alpha^{n-p}(1-\alpha)^{p-1}\,\phi(x_p)\,dx_p,$$

and mean
$$x_p = \frac{n!}{(n-p)!\,(p-1)!}\int_{-\infty}^{\infty} \alpha^{n-p}(1-\alpha)^{p-1}\,\phi(x_p)\,x_p\,dx_p$$
$$= \frac{n!}{(n-p)!\,(p-1)!}\int_{-\infty}^{\infty} \alpha^{n-p}(1-\alpha)^{p-1}\,x_p\,\frac{d\alpha}{dx_p}\,dx_p$$
$$= \frac{n!}{(n-p)!\,(p-1)!}\left\{-\int_0^{\infty}\frac{d}{dx_p}\left(\int_{x_p}^{\infty}(1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx_p}\,dx_p\right)x_p\,dx_p\right.$$
$$\left.+\int_{-\infty}^{0}\frac{d}{dx_p}\left(\int_{-\infty}^{x_p}(1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx_p}\,dx_p\right)x_p\,dx_p\right\}$$
$$= \frac{n!}{(n-p)!\,(p-1)!}\left\{\left[-\int_{x_p}^{\infty}(1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx_p}\,dx_p\,(x_p)\right]_0^{\infty}\right.$$
$$+\left[\int_{-\infty}^{x_p}(1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx_p}\,dx_p\,(x_p)\right]_{-\infty}^{0}$$
$$+\int_0^{\infty}dx_p\int_{x_p}^{\infty}(1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx_p}\,dx_p$$
$$\left.-\int_{-\infty}^{0}dx_p\int_{-\infty}^{x_p}(1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx_p}\,dx_p\right\}.$$

All the integrals in this expression are convergent (provided the frequency distribution is of finite range or behaves at ± ∞ like e^{-x^2}) and the terms in square brackets vanish at both limits.

Similarly we have:

$$\text{mean } x_q = \frac{n!}{(n-q)!\,(q-1)!}\left(\int_0^\infty dx_q \int_{x_q}^\infty (1-\alpha)^{q-1}\alpha^{n-q}\frac{d\alpha}{dx_q}\,dx_q - \int_{-\infty}^0 dx_q\int_{-\infty}^{x_q}(1-\alpha)^{q-1}\alpha^{n-q}\frac{d\alpha}{dx_q}\,dx_q\right).$$

Therefore mean $(x_p - x_q)$*

$$= \int_0^\infty \left\{\frac{\int_x^\infty (1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx}\,dx}{B(n-p+1,p)} - \frac{\int_x^\infty (1-\alpha)^{q-1}\alpha^{n-q}\frac{d\alpha}{dx}\,dx}{B(n-q+1,q)}\right\}dx$$

$$- \int_{-\infty}^0 \left\{\frac{\int_{-\infty}^x (1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx}\,dx}{B(n-p+1,p)} - \frac{\int_{-\infty}^x (1-\alpha)^{q-1}\alpha^{n-q}\frac{d\alpha}{dx}\,dx}{B(n-q+1,q)}\right\}dx$$

$$= \int_0^\infty \left\{1 - \frac{\int_{-\infty}^x (1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx}\,dx}{B(n-p+1,p)} - \left(1 - \frac{\int_{-\infty}^x (1-\alpha)^{q-1}\alpha^{n-q}\frac{d\alpha}{dx}\,dx}{B(n-q+1,q)}\right)\right\}dx$$

$$- \int_{-\infty}^0 \left\{\frac{\int_{-\infty}^x (1-\alpha)^{p-1}\alpha^{n-p}\frac{d\alpha}{dx}\,dx}{B(n-p+1,p)} - \frac{\int_{-\infty}^x (1-\alpha)^{q-1}\alpha^{n-q}\frac{d\alpha}{dx}\,dx}{B(n-q+1,q)}\right\}dx$$

$$= \int_{-\infty}^{\infty} \left\{\frac{\int_0^\alpha (1-\alpha)^{q-1}\alpha^{n-q}\,d\alpha}{B(n-q+1,q)} - \frac{\int_0^\alpha (1-\alpha)^{p-1}\alpha^{n-p}\,d\alpha}{B(n-p+1,p)}\right\}dx$$

$$= \int_{-\infty}^\infty \{I_\alpha(n-q+1,q) - I_\alpha(n-p+1,p)\}\,dx \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)},$$

---

* To prove mean $(x_p - x_q) = $ mean x_p − mean x_q we may proceed as follows:

Let δ be the difference between the values of the character in the pth and qth individuals in any *particular sample* of n individuals. Then

$$\delta = x_p - x_q.$$

This is true for every sample, therefore summing for all possible samples and dividing by their number

$$\bar{\delta} = \bar{x}_p - \bar{x}_q.$$

But to some minds it may seem better to approach the problem of this section by the same method as that used for consecutive individuals in the 1902 paper. If we do so we are easily led to the expression

$$\frac{n!}{(n-q)!\,(q-p-1)!\,(p-1)!}\int_{-\infty}^\infty\int_{-\infty}^{x_p} a'^{n-q}(a-a')^{q-p-1}(1-a)^{p-1}(x_p-x_q)\,\phi(x_p)\,\phi(x_q)\,dx_p\,dx_q,$$

for the mean difference, where

$$a = \int_{-\infty}^{x_p}\phi(x)\,dx, \quad a' = \int_{-\infty}^{x_q}\phi(x)\,dx.$$

By expanding $(a-a')^{q-p-1}$ and integrating by parts this may be shown to be equal to

$$\frac{n!}{(n-q)!\,(p-1)!}\int_{-\infty}^\infty \left[\sum_{s=0}^{s=q-p-1}\frac{(-1)^s a^{n-q+s+1}\int_a^1 a^{q-p-s-1}(1-a)^{p-1}\,da}{(q-p-s-1)!\,s!\,(n-q+s+1)}\right]dx,$$

where

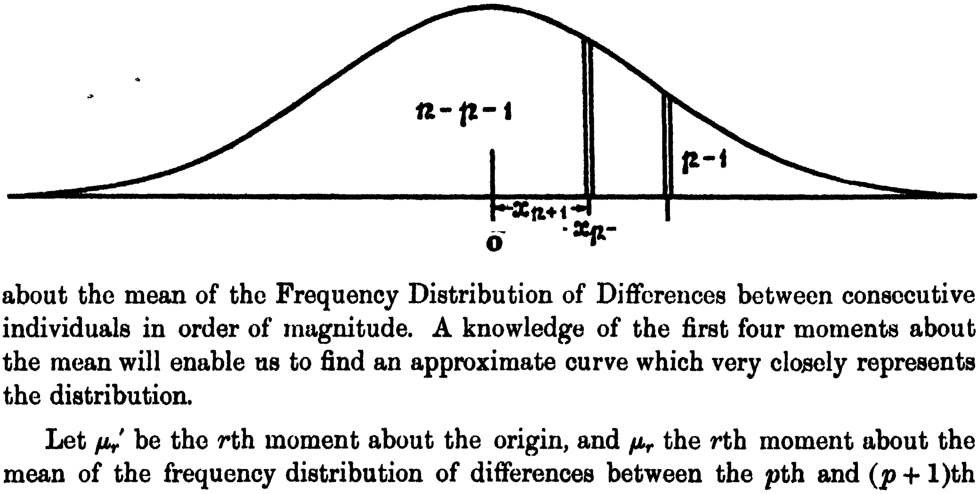
$$a = \int_{-\infty}^{x}\phi(x)\,dx.$$

where $I_x(r, s)$ denotes the ratio of the incomplete Beta-function $\int_0^x (1-x)^{r-1} x^{s-1}\, dx$ to the complete Beta-function $\int_0^1 (1-x)^{r-1} x^{s-1}\, dx$.

The numerical computation of this expression would involve the calculation of pairs of incomplete Beta-functions of α at convenient intervals of x and then performing a quadrature upon them. It would therefore be very troublesome, and it is questionable whether it would be any improvement on the process of finding the mean differences between the pth, $(p+1)$th, $(p+2)$th, ... qth individuals by formula (i) and summing the results.

(2) *Moments of the Frequency Distribution of Differences between Consecutive Individuals.*

We will now find expressions for the moments about any origin and thence



about the mean of the Frequency Distribution of Differences between consecutive individuals in order of magnitude. A knowledge of the first four moments about the mean will enable us to find an approximate curve which very closely represents the distribution.

Let μ_r' be the rth moment about the origin, and μ_r the rth moment about the mean of the frequency distribution of differences between the pth and $(p+1)$th individuals in random samples of n from the original population $[y = N\phi(x)]$.

Let α denote $\int_{-\infty}^{x_p} \phi(x)\, dx$ and α' denote $\int_{-\infty}^{x_{p+1}} \phi(x)\, dx$.

If we now expand the incomplete Beta-function and perform the summation for every term, the expression may be reduced to

$$\int_{-\infty}^{\infty} \left\{ \frac{n!}{(n-q+1)!\,(q-1)!}(1-\alpha)^{q-1}\alpha^{n-q+1} + \frac{n!}{(n-q+2)!\,(q-2)!}(1-\alpha)^{q-2}\alpha^{n-q+2} + \ldots + \frac{n!\,(1-\alpha)^p\alpha^{n-p}}{p!\,(n-p)!} \right\} dx$$

= mean difference between $(q-1)$th and qth individuals
+ mean difference between $(q-2)$th and $(q-1)$th individuals
+ mean difference between $(q-3)$rd and $(q-2)$th individuals
+ ....
+ mean difference between $(p-1)$th and pth individuals

a result which is easily seen to be equivalent to (iii) on expanding the two incomplete Beta-functions in that expression.

Then

$$\mu_r' = \frac{n!}{(n-p-1)!\,(p-1)!}\int_{-\infty}^{\infty}\int_{-\infty}^{x_p} \alpha'^{\,n-p-1}(1-\alpha)^{p-1}(x_p - x_{p+1})^r \frac{d\alpha}{dx_p}\frac{d\alpha'}{dx_{p+1}}\,dx_p\,dx_{p+1}$$

$$= \frac{n!}{(n-p-1)!\,(p-1)!}\int_{-\infty}^{\infty}(1-\alpha)^{p-1}\frac{d\alpha}{dx_p}\,dx_p\left\{\int_{-\infty}^{x_p}\alpha'^{\,n-p-1}(x_p - x_{p+1})^r\frac{d\alpha'}{dx_{p+1}}\,dx_{p+1}\right\}.$$

Now $$\int_{-\infty}^{x_p}\alpha'^{\,n-p-1}(x_p - x_{p+1})^r\frac{d\alpha'}{dx_{p+1}}\,dx_{p+1}$$

$$= \left[\frac{\alpha'^{\,n-p}}{n-p}(x_p - x_{p+1})^r\right]_{-\infty}^{x_p} + r\int_{-\infty}^{x_p}\frac{\alpha'^{\,n-p}}{n-p}(x_p - x_{p+1})^{r-1}\,dx_{p+1}.$$

The term in square brackets vanishes at both limits.

Therefore $$\mu_r' = \frac{r\,.\,n!}{(n-p)!\,(p-1)!}\int_{-\infty}^{\infty}(1-\alpha)^{p-1}\,Q\,\frac{d\alpha}{dx_p}\,dx_p,$$

where $$Q = \int_{-\infty}^{x_p}\alpha'^{\,n-p}(x_p - x_{p+1})^{r-1}\,dx_{p+1}.$$

Therefore $$\mu_r' = \frac{r\,.\,n!}{(n-p)!\,(p-1)!}\left\{\left[-\frac{(1-\alpha)^p}{p}\,Q\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty}\frac{(1-\alpha)^p}{p}\frac{dQ}{dx_p}\,dx_p\right\}.$$

The term in square brackets again vanishes at both limits and

$$\frac{dQ}{dx_p} = \int_{-\infty}^{x_p}\alpha'^{\,n-p}(r-1)(x_p - x_{p+1})^{r-2}\,dx_{p+1}.$$

Therefore $$\mu_r' = \frac{r(r-1)\,.\,n!}{(n-p)!\,p!}\int_{-\infty}^{\infty}\int_{-\infty}^{x_p}(1-\alpha)^p\,\alpha'^{\,n-p}(x_p - x_{p+1})^{r-2}\,dx_p\,dx_{p+1}$$

$$= \frac{r(r-1)\,.\,n!}{(n-p)!\,p!}\int_{-\infty}^{\infty}\int_{-\infty}^{x}(1-\alpha)^p\,\alpha'^{\,n-p}(x-y)^{r-2}\,dx\,dy,$$

where $$\alpha = \int_{-\infty}^{x}\phi(u)\,du,\quad \alpha' = \int_{-\infty}^{y}\phi(u)\,du \quad \text{..............................(iv).}$$

Putting $r = 2, 3, 4$ in succession we have:

$$\left.\begin{aligned}
\mu_2' &= \frac{2\,.\,n!}{(n-p)!\,p!}\int_{-\infty}^{\infty}\int_{-\infty}^{x}(1-\alpha)^p\,\alpha'^{\,n-p}\,dx\,dy\\
\mu_3' &= \frac{6\,.\,n!}{(n-p)!\,p!}\int_{-\infty}^{\infty}\int_{-\infty}^{x}(1-\alpha)^p\,\alpha'^{\,n-p}(x-y)\,dx\,dy\\
\mu_4' &= \frac{12\,.\,n!}{(n-p)!\,p!}\int_{-\infty}^{\infty}\int_{-\infty}^{x}(1-\alpha)^p\,\alpha'^{\,n-p}(x-y)^2\,dx\,dy\\
\text{and of course}\quad \mu_1' &= \frac{n!}{(n-p)!\,p!}\int_{-\infty}^{\infty}\alpha^{n-p}(1-\alpha)^p\,dx
\end{aligned}\right\}\text{............(v)}$$

and by the ordinary formulae for transferring moments to the mean

$$\left.\begin{aligned}
\mu_2 &= \mu_2' - \mu_1'^2\\
\mu_3 &= \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3\\
\mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4
\end{aligned}\right\}\text{ ....(vi).}$$

We have now to consider the best way of computing the expressions (v) numerically for particular values of r and p on the assumption that the original population is normally distributed.

We have to calculate integrals of the type

$$\int_{-\infty}^{\infty}\int_{-\infty}^{x} Z\,(x-y)^r\,dx\,dy,$$

where as many values of Z as desired can be calculated numerically, and r is either 0, 1 or 2.

By turning the axes through 45° this can be transformed into

$$\int_{0}^{\infty}\int_{-\infty}^{\infty} Z'\,dx'\,dy',$$

where $$Z' = Z\,(x-y)^r.$$

The integral is now in a shape in which it is easy to apply a cubature* formula and the process will be aided by the fact that $(x-y)$ is constant when x' is constant. We may note here that throughout all the numerical work in this paper, the standard deviation of the original population is taken as a unit.

The cubature formula we use is the following†:

$$\begin{aligned}\int_{0}^{\infty}\int_{-\infty}^{\infty} Z\,dx\,dy = \frac{hk}{120960}[&-191u_{-3}+1688u_{-2}-7843u_{-1}+60480u_0\\ &+128803u_1+119272u_2+121151u_3\\ &+120960\,(u_4+u_5+u_6+\dots)] \quad\dots\dots\dots\dots\dots\dots\dots(\text{vii}),\end{aligned}$$

where $u_r =$ sum of all the Z's at intervals k for which $x = rh$. This formula is correct up to and including seventh order differences.

The process, in this case, amounts to calculating ordinates at convenient equal distances, summing them in diagonals and multiplying by the appropriate coefficients. For example, in the case $n = 10$, $p = 1$, values of $(1-\alpha)\,\alpha'^9$ are calculated at intervals of quarter units of standard deviation from $x = -1{\cdot}75$ to $5{\cdot}00$ and $y = -1{\cdot}00$ to $x + {\cdot}75$ (for smaller values of x and y the ordinates vanish to seven places of decimals).

If we require the second moment of the frequency distribution of differences between the first and second individuals in samples of 10, we have only to sum the successive diagonals beginning from the left to obtain u_{-3}, u_{-2}, u_{-1}, etc.

If we require the third moment the u's are obtained by multiplying the sums of the successive diagonals by -3, -2, -1, 0, 1, 2, 3, etc., and dividing each by 4, while for the fourth moment the appropriate multipliers are $\frac{9}{16}$, $\frac{4}{16}$, 0, $\frac{1}{16}$, $\frac{4}{16}$, $\frac{9}{16}$, etc. The results obtained by the cubature have then to be multiplied by the appropriate factorial terms given in (v) (for the case of $n = 10$, $p = 1$, these will be 20, 60, 120)‡.

---

* Cubature is the name sometimes given to the process of calculation of volumes which is analogous to quadrature for areas.

† A proof of this formula will be found in *Tracts for Computers*, No. x. p. 27.

‡ The numerical work for this case is given fully in *Tracts for Computers*, No. x. p. 30.

In this manner we have calculated the first four moments of the frequency distribution of differences for the values $n = 3, 10, 50, 100, 500, 1000$ and $p = 1$, $p = 2$ on the assumption that the character is normally distributed in the original population.

We find:

$p = 1$

n	3	10	50	100	500	1000
$\mu_1' = \chi_{n,p}$	·8458	·5388	·3942	·3594	·3044	·2873
μ_2'	1·1730	·5054	·2830	·2378	·1735	·1554
μ_3'	2·1158	·6424	·2825	·2201	·1401	·1196
μ_4'	4·5575	1·0004	·3520	·2558	·1432	·1171

$p = 2$

n	3	10	50	100	500	1000
$\mu_1' = \chi_{n,p}$	·8458	·3453	·2263	·2018	·1655	·1550
μ_2'	1·1730	·2155	·0961	·0772	·0526	·0463
μ_3'	2·1158	·1858	·0581	·0423	·0241	·0201
μ_4'	4·5575	·1990	·0437	·0292	·0141	·0112

while the moments about the mean and β_1, β_2 are as follows:

$p = 1$

n	3	10	50	100	500	1000
$\mu_2 = \sigma_d^2$	·4577	·2151	·1276	·1086	·0808	·0729
μ_3	·3495	·1383	·0704	·0565	·0381	·0331
μ_4	·8988	·2434	·0980	·0736	·0433	·0363
β_1	1·28	1·92	2·39	2·49	2·75	2·83
β_2	4·29	5·26	6·03	6·24	6·63	6·83

$p = 2$

n	3	10	50	100	500	1000
$\mu_2 = \sigma_d^2$	·4577	·0963	·0449	·0365	·0252	·0223
μ_3	·3495	·0449	·0161	·0120	·0071	·0060
μ_4	·8988	·0540	·0127	·0090	·0044	·0037
β_1	1·28	2·26	2·86	2·96	3·15	3·25
β_2	4·29	5·82	6·30	6·76	6·93	7·44

It may also be of interest to give the standard deviations and coefficients of variation $\left(100\frac{\sigma_d}{\mu_1'}\right)$ in the twelve cases.

p :

n	*3*	*10*	*50*	*100*	*500*	*1000*
σ_d $100\sigma_d/\mu_1'$	·6765 79·98	·4638 86·08	·3572 90·61	·3295 91·68	·2843 93·40	·2699 93·94

$p = 2$

n	*3*	*10*	*50*	*100*	*500*	*1000*
σ_d $100\sigma_d/\mu_1'$	·6765 79·98	·3103 89·86	·2119 93·64	·1910 94·65	·1588 95·95	·1494 96·39

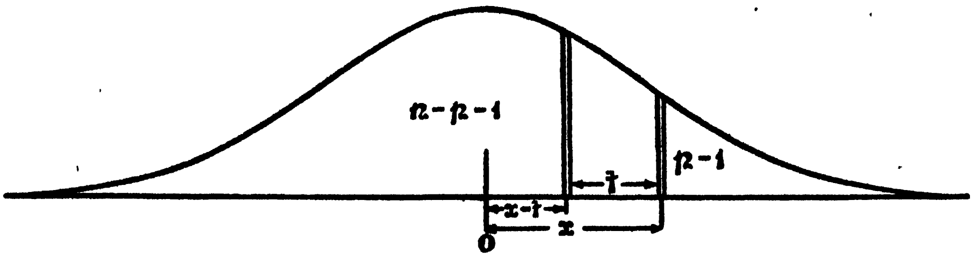
It will be seen that both for $p = 1$ and $p = 2$, although the variability decreases fairly rapidly, the coefficient of variation increases steadily and seems to tend to 100 °/ₒ as n tends to infinity.

The moments above calculated will be used to obtain curves which approximately represent the frequency distribution of differences.

But before obtaining approximations it will be well to consider more closely their exact nature. It is the purpose of the next section to determine the actual as distinguished from the approximate frequency distributions and where possible to submit them to numerical calculation. Where the latter is not possible our investigation may nevertheless assist us in determining what type of curve may reasonably be used as an approximation.

(3) *Actual Frequency Distribution of Differences between Consecutive Individuals.*

The chance of the pth and $(p+1)$th individuals in descending order of magnitude (in a sample of n) having a difference between t and $t+dt$ is, on the assumption that the original population is normally distributed,



$$\frac{n!}{(n-p-1)!\,(p-1)!}\int_{x=-\infty}^{x=\infty}\alpha'^{\,n-p-1}(1-\alpha)^{p-1}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-t)^2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\right)dx\,.\,dt$$

$$=\frac{n!}{(n-p-1)!\,(p-1)!}\int_{x=-\infty}^{x=\infty}\alpha'^{\,n-p-1}(1-\alpha)^{p-1}\frac{1}{2\pi}e^{-\{(x-\frac{1}{2}t)^2+\frac{1}{4}t^2\}}\,dx\,.\,dt,$$

where $$\alpha' = \int_{-\infty}^{x-t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du, \quad \alpha = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du^{*} \quad \ldots\ldots\ldots\ldots\text{(viii)}.$$

Putting $x - \frac{1}{2}t = x'$ we obtain for the above expression

$$\frac{n!}{(n-p-1)!\,(p-1)!} \int_{x'=-\infty}^{x'=\infty} \alpha'^{\,n-p-1} (1-\alpha)^{p-1} \frac{1}{2\pi} e^{-x'^2} e^{-\frac{1}{4}t^2} dx'.dt \;\ldots\ldots\text{(ix)}.$$

Now $$\alpha' = \int_{-\infty}^{x'-\frac{1}{2}t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du, \quad \alpha = \int_{-\infty}^{x'+\frac{1}{2}t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du.$$

Therefore $$\frac{\partial\alpha'}{\partial t} = -\frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x'-\frac{1}{2}t)^2} = -\frac{1}{2}\frac{\partial\alpha'}{\partial x'},$$

$$\frac{\partial\alpha}{\partial t} = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x'+\frac{1}{2}t)^2} = \frac{1}{2}\frac{\partial\alpha}{\partial x'}.$$

Now if we expand α, α' in powers of t we obtain:

$$\alpha = \alpha_0 + \left(\frac{\partial\alpha}{\partial t}\right)_0 t + \left(\frac{\partial^2\alpha}{\partial t^2}\right)_0 \frac{t^2}{2!} + \ldots + \left(\frac{\partial^r\alpha}{\partial t^r}\right)_0 \frac{t^r}{r!} + \ldots,$$

$$\alpha' = \alpha_0' + \left(\frac{\partial\alpha'}{\partial t}\right)_0 t + \left(\frac{\partial^2\alpha'}{\partial t^2}\right)_0 \frac{t^2}{2!} + \ldots + \left(\frac{\partial^r\alpha'}{\partial t^r}\right)_0 \frac{t^r}{r!} + \ldots,$$

where the suffix 0 denotes that t is to be put zero after the differentiation.

But $$\alpha_0 = \int_{-\infty}^{x'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du,$$

$$\left(\frac{\partial\alpha}{\partial t}\right)_0 = \frac{1}{2}\frac{\partial}{\partial x'}\int_{-\infty}^{x'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \frac{1}{2}\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2},$$

$$\left(\frac{\partial^r\alpha}{\partial t^r}\right)_0 = \left(\frac{1}{2}\right)^r \frac{d^{r-1}}{dx'^{r-1}}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right);$$

Similarly $$\left(\frac{\partial^r\alpha'}{\partial t^r}\right)_0 = \left(-\frac{1}{2}\right)^r \frac{d^{r-1}}{dx'^{r-1}}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right).$$

Therefore

$$\alpha = \int_{-\infty}^{x'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du + \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right)\frac{t}{2} + \frac{1}{2!}\frac{d}{dx'}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right)\left(\frac{t}{2}\right)^2 + \ldots$$
$$+ \frac{1}{r!}\frac{d^{r-1}}{dx'^{r-1}}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right)\left(\frac{t}{2}\right)^r + \ldots \;\ldots\ldots\ldots\ldots\ldots\text{(x)},$$

$$\alpha' = \int_{-\infty}^{x'} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du - \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right)\frac{t}{2} + \frac{1}{2!}\frac{d}{dx'}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right)\left(\frac{t}{2}\right)^2 - \ldots$$
$$+ \frac{(-1)^r}{r!}\frac{d^{r-1}}{dx'^{r-1}}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}\right)\left(\frac{t}{2}\right)^r + \ldots \;\ldots\ldots\ldots\text{(x } bis\text{)}.$$

If now we substitute the expressions (x) and (x *bis*) for α and α' in (ix), re-expand and integrate term by term, we obtain the frequency distribution of differences in the form

$$e^{-\frac{1}{4}t^2} \text{ (Series in Ascending Powers of } t\text{)}.$$

---

* We take as our unit of measurement the standard deviation of the original population.

A mathematical expression for the general term may be found by the use of Burmann's Theorem* for expanding one function in powers of another.

Burmann's Theorem may be written

$$f(z)=f(a)+\sum_{r=1}^{\infty}\frac{B_r[F(z)]^r}{r!},$$

where
$$B_r=\left[\frac{d^{r-1}}{dz^{r-1}}\left\{\left(\frac{z-a}{F(z)}\right)^r f'(z)\right\}\right]_{z=a},$$

and a is a solution of the equation $F(z)=0$.

Let
$$t=F(\alpha')=\phi(1-\alpha),$$

i.e. the functions $F(\alpha')$ and $\phi(1-\alpha)$ are the inverse functions of

$$\alpha'=\int_{-\infty}^{x'-\frac{1}{2}t}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du \quad\text{and}\quad 1-\alpha=\int_{x'+\frac{1}{2}t}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du\dagger.$$

Then a is given by
$$F(\alpha')=\phi(\alpha)=0,$$

or
$$a=\int_{-\infty}^{x'}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du.$$

Accordingly
$$\alpha'^{n-p-1}=a^{n-p-1}+\sum_{r=1}^{\infty}\frac{B_r t^r}{r!}$$

where
$$B_r=\left[\frac{d^{r-1}}{d\alpha'^{r-1}}\left\{\left(\frac{\alpha'-a}{t}\right)^r(n-p-1)\,\alpha'^{n-p-2}\right\}\right]_{\substack{\alpha'=a,\\ \text{i.e. } t=0}} \qquad \text{............(xi).}$$

Similarly
$$(1-\alpha)^{p-1}=(1-a)^{p-1}+\sum_{r=1}^{\infty}\frac{C_r t^r}{r!},$$

where
$$C_r=-\left[\frac{d^{r-1}}{d\alpha^{r-1}}\left\{\left(\frac{\alpha-a}{t}\right)^r(p-1)(1-\alpha)^{p-2}\right\}\right]_{\substack{\alpha=a,\\ \text{i.e. } t=0}} \qquad \text{.........(xi \textit{bis}).}$$

Therefore
$$\alpha'^{n-p-1}(1-\alpha)^{p-1}=U_0+\sum_{r=1}^{\infty}\frac{U_r t^r}{r!},$$

where
$$U_r=B_0C_r+rB_1C_{r-1}+\frac{r(r-1)}{2!}B_2C_{r-2}+\ldots+B_rC_0 \qquad \text{.........(xii).}$$

Therefore the frequency distribution is

$$y=\frac{n!}{(n-p-1)!\,(p-1)!}\,\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{4}t^2}\sum_{0}^{\infty}\frac{V_r t^r}{r!} \qquad \text{...............(xiii)}$$

where
$$V_r=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}U_r e^{-x'^2}dx'$$

and U_r, B_r, C_r have the above values; also

$$B_0=a^{n-p-1}=\left\{\int_{-\infty}^{x'}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du\right\}^{n-p-1},$$

$$C_0=(1-a)^{p-1}=\left\{\int_{x'}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du\right\}^{p-1},$$

$$U_0=B_0C_0.$$

---

* See Edwards' *Differential Calculus*, 1912 edition, p. 455 and Whittaker and Watson's *Modern Analysis*, p. 129.

† These inverse functions certainly exist, since for a given x', there is only one value of a corresponding to one of t and vice versa.

We thus have a formal solution to the problem; but even if it could be shown theoretically that the series was always convergent, it would still have to be admitted that the form is not easily adaptable to numerical calculation.

The calculation of the B's and C's in the first place involves the evaluation of a number of undetermined forms, which becomes somewhat troublesome when r is greater than 3. In practice, it is easier to obtain as many terms as desired in the expansions of α'^{n-p-1} and $(1-\alpha)^{p-1}$ direct from the expansions for α' and $1-\alpha$, by Arbogast's rule or some other convenient process*. We have, however, succeeded in working out completely the cases of $n=2$ and $n=3$.

(α) $n=2$, $p=1$. In this case $B_0=1$, $C_0=1$, $U_0=1$, $B_r=U_r=C_r=0$ for all positive integral values of r,

$$V_r=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x'^2}dx'=\frac{1}{\sqrt{2}},$$

and the distribution is

$$y=\sqrt{2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{4}t^2}=\frac{1}{\sqrt{\pi}}e^{-\frac{1}{4}t^2} \quad \text{.....................(xiv)†}.$$

The range is from 0 to ∞, since the first individual has, by hypothesis, a greater value of the character than the second, and so the differences will always be of the same sign which we take to be positive.

Thus we have the interesting result that the frequency distribution of differences between pairs of individuals taken at random from a normal population, is itself a normal curve whose standard deviation is $\sqrt{2}$ times the standard deviation of the original population.

(β) $n=3, p=1$. Here $B_0=\alpha=\int_{-\infty}^{x'}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du$, $C_0=1$,

$$\alpha'^{n-p-1}=\alpha',\quad (1-\alpha)^{p-1}=1.$$

Therefore $$U_r=B_r=\frac{1}{2^r}(-1)^r\frac{d^{r-1}}{dx'^{r-1}}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x'^2}\right),\quad C_r=0,$$

for all positive integral values of r.

---

* Arbogast's rule in its most convenient form is as follows: if

$$\left(a_0+a_1x+a_2\frac{x^2}{2!}+a_3\frac{x^3}{3!}+\dots\right)^n$$

be expanded in the form

$$A_0+A_1x+A_2\frac{x^2}{2!}+A_3\frac{x^3}{3!}+\dots$$

the coefficients $A_1, A_2, A_3, \dots$ can be successively determined from

$$A_{r+1}=\left(a_1\frac{\partial}{\partial a_0}+a_2\frac{\partial}{\partial a_1}+a_3\frac{\partial}{\partial a_2}+\dots\right)A_r.$$

See Edwards' *Differential Calculus*, p. 450.

† The moments of this distribution are easily found and are as follows:

$$\mu_1'=\frac{2}{\sqrt{\pi}},\quad \mu_2'=2,\quad \mu_3'=\frac{8}{\sqrt{\pi}},\quad \mu_4'=12,\quad \mu_2=2\left(1-\frac{2}{\pi}\right),\quad \mu_3=\frac{4}{\sqrt{\pi}}\left(\frac{4}{\pi}-1\right),\quad \mu_4=4\left(3-\frac{4}{\pi}-\frac{12}{\pi^2}\right).$$

This value of μ_1' is in agreement numerically with that found by Pearson in the 1902 paper.

By the method of Section (2) we had:

$$\text{area} = 1{\cdot}0000, \quad \mu_1' = {\cdot}8458, \quad \mu_2' = 1{\cdot}1730, \quad \mu_3' = 2{\cdot}1158, \quad \mu_4' = 4{\cdot}5575.$$

The agreement seems satisfactory, as the divergence in μ_3' and μ_4' is due to stopping at the coefficient of t^{17} and neglecting higher powers in our original series (xviii). The moments about the mean by the two methods are given below:

	μ_2	μ_3	μ_4
Method of Section 2	·4565	·3656	·8629
Actual Series	·4577	·3495	·8988

The divergences in μ_3 and μ_4 occur for the same reason as in μ_3' and μ_4'.

We have further calculated from (xviii) the ordinates of the curve at intervals of 0·5, so that it may be plotted.

t	y	t	y
0·0	·84628	3·0	·01958
0·5	·66642	3·5	·00613
1·0	·45022	4·0	·00168
1·5	·26055	4·5	·00044
2·0	·12898	5·0	·00011
2·5	·05458	5·5	·00003
		6·0	·00001

(γ) $n = 10, p = 1$. An attempt, extending over some weeks, was made to compute the actual frequency distribution for $n = 10, p = 1$.

In this case it may be seen from (ix) and (xiii) that the frequency distribution is:

$$y = \frac{90}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} \sum_0^\infty \frac{V_r t^r}{r!},$$

where

$$V_r = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} B_r e^{-x'^2} dx' \quad \text{.........................(xx)},$$

and $\frac{B_r}{r!}$ is the coefficient of t^r in the expansion of α'^8 in ascending powers of t, α' being given by (x *bis*). The first ten terms in this expansion were worked out by Arbogast's theorem and the first eight V's were actually computed; but the resulting series was not sufficiently convergent to make the numerical computation of the frequency distribution possible from a knowledge of these eight terms. The attempt was then abandoned owing to the amount of arithmetic involved in the computation of the higher V's becoming prohibitive. To show that this is so we need only mention that B_8 is given by

$$\begin{aligned} B_8 = \tfrac{1}{32} [&5040z^8 + 141120az^6z_1 + 2520a^2(56z^5z_2 + 210z^4z_1^2) \\ &+ 4200a^3(14z^4z_3 + 112z^3z_1z_2 + 84z^2z_1^3) \\ &+ 210a^4(56z^3z_4 + 420z^2z_1z_3 + 280z^2z_2^2 + 840zz_1^2z_2 + 105z_1^4) \\ &+ 42a^5(28z^2z_5 + 168zz_1z_4 + 280zz_2z_3 + 280z_1z_2^2 + 210z_1^2z_3) \\ &+ 7a^6(8zz_6 + 28z_1z_5 + 56z_2z_4 + 35z_3^2) + a^7z_7] \quad \text{.......................(xxi)}, \end{aligned}$$

where
$$a=\int_{-\infty}^{x'}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du,\quad z=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x'^2},$$
and
$$z_r=\frac{d^r}{dx'^r}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x'^2}\right).$$

Sufficient numerical values of B_8 have to be calculated in order to make it possible to obtain V_8 by a quadrature; further the number of terms in B_r increases with r.

It would thus seem that for the higher values of n we are thrown back on the method of fitting curves to the moments calculated in Section (2) in order to obtain information about the frequency distributions.

This information may however be supplemented in one respect, by the computation of the values of y in the actual curves at the start of their range, i.e. when $t=0$.

We then have
$$y=\frac{n!}{(n-p-1)!\,(p-1)!}\int_{x'=-\infty}^{x'=\infty}a^{n-p-1}(1-a)^{p-1}z^2dx',$$
where
$$a=\int_{-\infty}^{x'}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du,\quad z=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x'^2}\ \ldots\ldots\ldots\ldots\ldots\ldots\text{(xxii)}.$$

For $p=1$, $n=10, 50, 100, 500, 1000$ respectively, this reduces to
$$90\int_{-\infty}^{\infty}a^8z^2dx',$$
$$2450\int_{-\infty}^{\infty}a^{48}z^2dx',$$
$$9900\int_{-\infty}^{\infty}a^{98}z^2dx',$$
$$249{,}500\int_{-\infty}^{\infty}a^{498}z^2dx',$$
$$999{,}000\int_{-\infty}^{\infty}a^{998}z^2dx',$$
and for $p=2$, $n=10, 50, 100, 500, 1000$ to
$$720\int_{-\infty}^{\infty}a^7(1-a)\,z^2dx',$$
$$117{,}600\int_{-\infty}^{\infty}a^{47}(1-a)\,z^2dx',$$
$$970{,}200\int_{-\infty}^{\infty}a^{97}(1-a)\,z^2dx',$$
$$12{,}425{,}100\int_{-\infty}^{\infty}a^{497}(1-a)\,z^2dx',$$
$$997{,}002{,}000\int_{-\infty}^{\infty}a^{997}(1-a)\,z^2dx'.$$

These integrals may be calculated by means of a quadrature. This has been done, taking ordinates at intervals of ·2, it being necessary for $n=100$, 500 and

1000 to take "a" to ten places and use ten figure logarithms. We have the following results:

Values of Ordinate when $t = 0$.

	$n=10$	$n=50$	$n=100$	$n=500$	$n=1000$
$p=1$	1·539	2·249	2·495	3·037	3·241
$p=2$	2·537	4·104	4·656	5·769	6·193

(4) *Approximations to the Frequency Distribution of Differences between Consecutive Individuals.*

We will now turn to the problem of finding approximations to the frequency distributions by fitting curves to the moments calculated in Section (2). Much difficulty has been experienced in finding a curve of a comparatively simple type which is at the same time an adequate representation of the distribution for all the cases that have been considered. There are of course only two cases in which the adequacy of the approximation may be directly tested, namely, $n = 2$ and $n = 3$, for these are the only two cases in which the ordinates may be calculated. In other cases we must be content to compare the first four moments about the mean and the initial ordinates of the actual and approximate curves. This is, however, adequate, as if these five quantities are in good agreement, the fit may be regarded as good.

Three types of curve suggest themselves as possibilities for approximation.

(α) A Pearson frequency-curve.

(β) A curve of the form $e^{-\frac{1}{2}x^2}$(polynomial in x), i.e. the sum of a *finite number* of tetrachoric functions.

(γ) The tail of a normal curve, i.e. a curve of the form

$$y = y_0 e^{-\frac{1}{2}\left\{\frac{(x+h)^2 - h^2}{\Sigma^2}\right\}}.$$

(α) Turning to the values of β_1, β_2 given on page 107, we see that they lead to the following types:

$$\left.\begin{array}{l} p=1\left\{\begin{array}{l} n=3\\ n=10\\ n=50\\ n=100\\ n=500\\ n=1000\end{array}\right. \\ p=2\left\{\begin{array}{l} n=3\\ n=10\end{array}\right.\end{array}\right\}$$

Type I_I,

i.e. $y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$, where the origin is at the mode and a_1, a_2, m_1, m_2 are positive.

$$p=2\left\{\begin{array}{l} n=50\\ n=100\\ n=500\\ n=1000\end{array}\right\}$$

Type J_I,

i.e. $y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\Big/\left(\frac{x}{a_2}-1\right)^{m_2}$, where the origin is at $x = 0$ and a_1, a_2, m_1, m_2 are positive.

Now these curves start in the one case with a zero ordinate, and in the other case with an infinite ordinate and so are not adapted to the representation of the actual distributions which we know start with a finite ordinate. If their constants be worked out from the moments, the curves may be shown to fit well except near the beginning of the range, but their failure there is a grave disadvantage.

It may be noticed if we examine the β_1, β_2 plane that in all the cases considered the point β_1, β_2 lies near that portion of the biquadratic curve along which the appropriate type is Pearson's Type IX, i.e. $y = y_0\left(1 - \frac{x}{a}\right)^m$. This is a J curve starting with a finite ordinate. It has accordingly been tried for the case of $n = 3$, a and m being determined from the known values of y_0 and μ_1', but the fit while naturally good at the beginning of the range is on the whole worse than before.

This method has therefore to be abandoned.

(β) The next method which suggests itself is to take a curve of the form $e^{-\frac{1}{2}x^2}(a + bx + cx^2 + dx^3)$ or of the form $e^{-\frac{1}{2}x^2}(a + bx + cx^2 + dx^3 + ex^4 + fx^5)$ and to determine the constants from the known values of the initial ordinate, the area and in one case the first two and in the other case the first four moments about $x = 0$.

Now for $n = 3$ both these forms lead to an excellent fit but when the method is tried for higher values of n it gives rise to negative frequencies for large values of x; showing that 3 or 5 terms are not enough adequately to represent what is really an infinite series, even when the coefficients of the terms retained are modified so as to make the moments agree with their actual values. So that this method too is in general inapplicable.

(γ) The fitting of a curve of the form

$$y = y_0 e^{-\frac{1}{2}\left\{\frac{(x+h)^2 - h^2}{\Sigma^2}\right\}}$$

is perhaps suggested by the fact that when $n = 2$ the frequency distribution of differences is accurately of this type and further that it would seem to be possible to expand y in the form $e^{-\frac{1}{2}x^2}$ (Series in powers of x), a form which has actually been found to represent the distribution.

This curve is fitted by making its mean and standard deviation identical with the actual mean and standard deviation, which leads to equations from which y_0, h and Σ may be determined. A comparison between the actual values of y_0, μ_3 and μ_4 and those given by the approximate curve is then a test of the adequacy of the fit in those cases when the ordinates of the actual distribution cannot be directly calculated.

It has been shown by Pearson and Lee[*] that if $h' = \frac{h}{\Sigma}$, and if $m_r(h')$, or for brevity m_r', represents the incomplete normal moment function $\int_{h'}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} x^r dx$, and that if $y_0 = \frac{M}{\Sigma}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h'^2}\right)$, so that M is the area of the complete normal curve of which our curve is a portion, then

---

[*] See *Biometrika*, Vol. VI. p. 68.

$$\left(\frac{\sigma_d}{\mu_1'}\right)^2 = \psi_1 = \frac{\mu_2' - \mu_1'^2}{\mu_1'^2} = \frac{m_2' m_0' - m_1'^2}{(m_1' - h'm_0')^2},$$

$$M = 1/m_0'^{*},$$

$$\Sigma = \mu_1' m_0'/(m_1' - h'm_0') = \mu_1'\psi_2 \text{..............(xxiii).}$$

ψ_1 and ψ_2 have been tabled against h'. Hence we may find ψ_1 from the actual mean and standard deviation and given ψ_1 we may find h' from the table. Then from h' ψ_2 is found and accordingly:

$$\Sigma = \mu_1'\psi_2,$$

and finally

$$y_0 = \frac{M}{\Sigma} z_h' \text{ ....(xxiv).}$$

To test the adequacy of the fit μ_3 and μ_4 can then be calculated from the relations

$$\mu_3 = \Sigma^3 \left\{\frac{m_3'}{m_0'} - 3\frac{m_1'}{m_0'}.\frac{m_2'}{m_0'} + 2\left(\frac{m_1'}{m_0'}\right)^3\right\},$$

$$\mu_4 = \Sigma^4 \left\{\frac{m_4'}{m_0'} - 4\frac{m_3'}{m_0'}.\frac{m_1'}{m_0'} + 6\frac{m_2'}{m_0'}.\left(\frac{m_1'}{m_0'}\right)^2 - 3\left(\frac{m_1'}{m_0'}\right)^4\right\} \text{ ....(xxv),}$$

and together with y_0 compared with their actual values.

If

$$y = y_0 e^{-\frac{1}{2}\left\{\frac{(x+h)^2 - h^2}{\Sigma^2}\right\}} \text{ ....(xxvi)}$$

be fitted in this manner by using the means and standard deviations given on p. 107, we find the following values for y_0, h and Σ:

$p=1$				$p=2$			
n	h	Σ	y_0	n	h	Σ	y_0
2	0	1·414	·564	*2*	—	—	—
3	·525	1·271	·848	*3*	·525	1·271	·848
10	1·295	1·095	1·529	*10*	1·609	·878	2·540
50	2·133	1·060	2·250	*50*	2·295	·785	4·105
100	2·384	1·046	2·508	*100*	2·587	·774	4·655
500	2·896	1·027	3·037	*500*	3·071	·749	5·761
1000	3·116	1·025	3·237	*1000*	3·170	·733	6·148
				1200	3·189	·729	6·289

These values if substituted in (xxvi) give us the approximate frequency distributions of differences between consecutive individuals.

To test the adequacy of the approximations we have the following table comparing the actual values of y_0, μ_3 and μ_4 with the values obtained from (xxiv) and (xxv).

*** Here we are supposing the area of the tail to be unity so that the total frequency of our approximate distribution will be unity as it was taken to be in the actual distribution of Section (8). If M' be the area of the tail we have $M = M'/m_0'$.**

$p = 1$

	$n=3$		$n=10$		$n=50$		$n=100$		$n=500$		$n=1000$	
	Actual	Fitted	Actual	Fitted	Actual	Fitted	Actual	Fitted	Actual	Fitted	Actual	Fitted
y_0	·848	·848	1·539	1·529	2·249	2·250	2·495	2·508	3·037	3·037	3·241	3·237
μ_3	·350	·352	·138	·136	·070	·069	·057	·057	·038	·036	·033	·034
μ_4	·899	·906	·243	·239	·098	·104	·074	·070	·043	—*	·036	—

$p = 2$

	$n=3$		$n=10$		$n=50$		$n=100$		$n=500$		$n=1000$	
	Actual	Fitted	Actual	Fitted	Actual	Fitted	Actual	Fitted	Actual	Fitted	Actual	Fitted
y_0	·848	·848	2·537	2·540	4·104	4·105	4·656	4·660	5·769	5·761	6·193	6·148
μ_3	·350	·352	·045	·045	·016	·015	·012	·008	·007	·007	·006	·006
μ_4	·899	·906	·054	·055	·013	·017	·009	—	·004	—	·004	—

This shows a satisfactory agreement.

In the case of $n = 3, p = 1$ or 2 we may proceed further and compare the ordinates of the actual frequency distribution (xviii) with those of its approximation

$$y = \cdot 848\, e^{-\frac{1}{2}\left\{\frac{(x+\cdot 525)^2 - (\cdot 525)^2}{(1\cdot 271)^2}\right\}}.$$

We find:

x	y	
	Actual	Fitted
0	·846	·848
·5	·666	·667
1·0	·450	·449
1·5	·261	·260
2·0	·129	·128
2·5	·055	·054
3·0	·020	·020
3·5	·006	·006
4·0	·002	·002
4·5	·000	·000

which also shows a satisfactory agreement.

---

* The values of μ_4 in the approximate curves (when $n = 500$ and 1000, $p = 1$ and $n = 100$, 500 and 1000, $p = 2$) cannot be obtained because three decimal places in h are inadequate for this purpose and for such large values of h we cannot table ψ_1 to more than three figures even using Sheppard's Tables of the Probability Integral to twelve figures. Therefore h itself cannot be obtained to more figures.

The approximate distributions are shown in Diagrams 1 and 2, while in Diagrams 3, 4, 5 h, Σ and y_0 are plotted against n. The points lie on smooth curves and thus it is possible to obtain from these graphs the approximate frequency distributions of differences between consecutive individuals for other values of n when p is either 1 or 2; that is the approximate frequency distribution of differences between the first and second or second and third individuals in samples of any size.

In Diagram 1 the unit of the horizontal scale is the standard deviation of the original population, the total frequency is taken as unity, and where the reading of the vertical scale is y the corresponding frequency is $y\delta x$. The standard deviations in Diagram 3 and the h's in Diagrams 4 and 5 are also measured in terms of the standard deviation of the original population as a unit.

CONCLUSION.

(1) We have shown that the frequency distributions of differences between the first and second and between the second and third individuals in order of magnitude, in samples of any size taken from a normal population, are closely represented by J curves whose equations may be written in the form

$$y = y_0 e^{-\frac{1}{2}\left\{\frac{(x+h)^2 - h^2}{\Sigma^2}\right\}}.$$

We will now consider briefly what is the bearing of this on Galton's prize problem. Galton said that the proper ratio in which to divide the prize was the ratio of the mean excess in merit of the first individual over the second to that of the second over the third.

But it may be urged that we should take not the means but the most probable values of these quantities and find their ratio. This would lead us to a difficulty, for each of these is zero*, or we have to evaluate in some way an undetermined form, and it is not clear how this should be done. For our own part, we should be disposed to keep to the means though fully realising that this is somewhat arbitrary†.

(2) We will now consider how to obtain the correlation between brothers by averaging

(α) The differences, all taken positively, of pairs of individuals drawn at random from the population.

(β) The differences, all taken positively, of pairs of brothers drawn at random.

We shall assume that the original population is normally distributed and that the bivariate correlation surface for pairs of brothers is also normal.

---

* The idea that the most probable value should occur when the two individuals are exactly alike ceases to be strange when it is realised that the probability of any *particular* value is infinitesimal.

† [The writer we think overlooks what Galton was aiming at; he was not seeking the most probable individual result, but how to be just *on the average* to competitors. ED.]

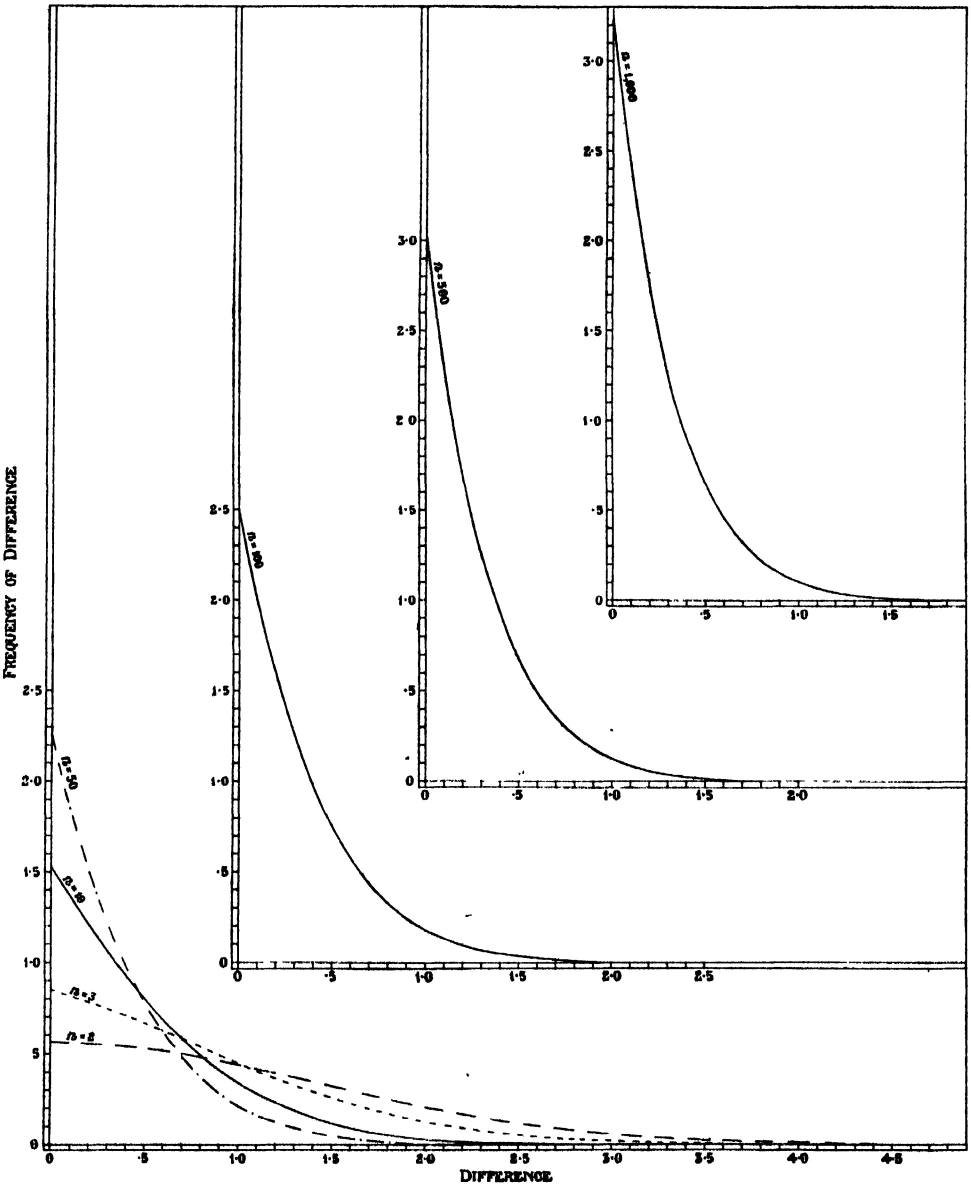

Approximate Frequency Distributions
of Differences Between First & Second Individuals

Diagram 1

APPROXIMATE FREQUENCY DISTRIBUTIONS
OF DIFFERENCES BETWEEN SECOND & THIRD INDIVIDUALS

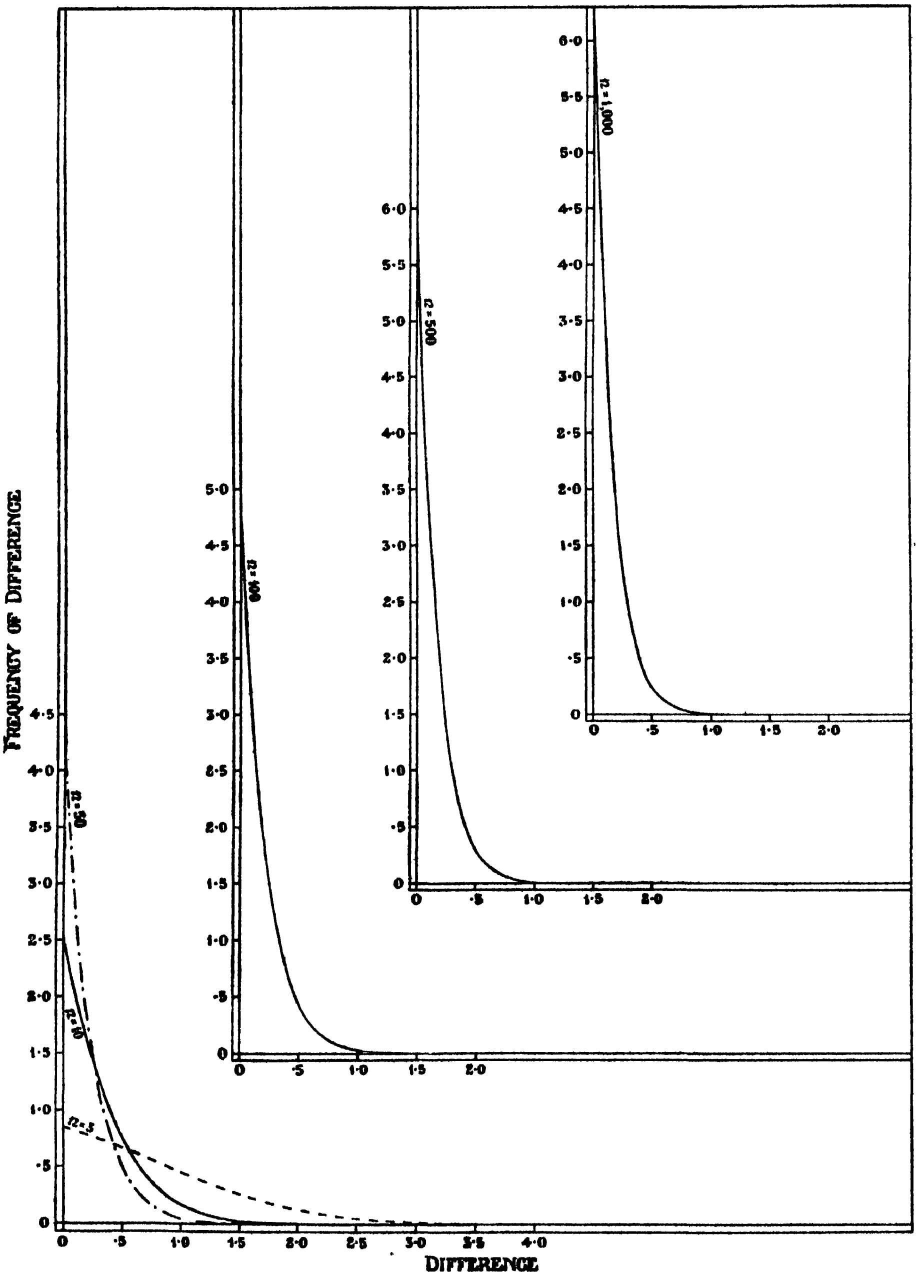


Diagram 2

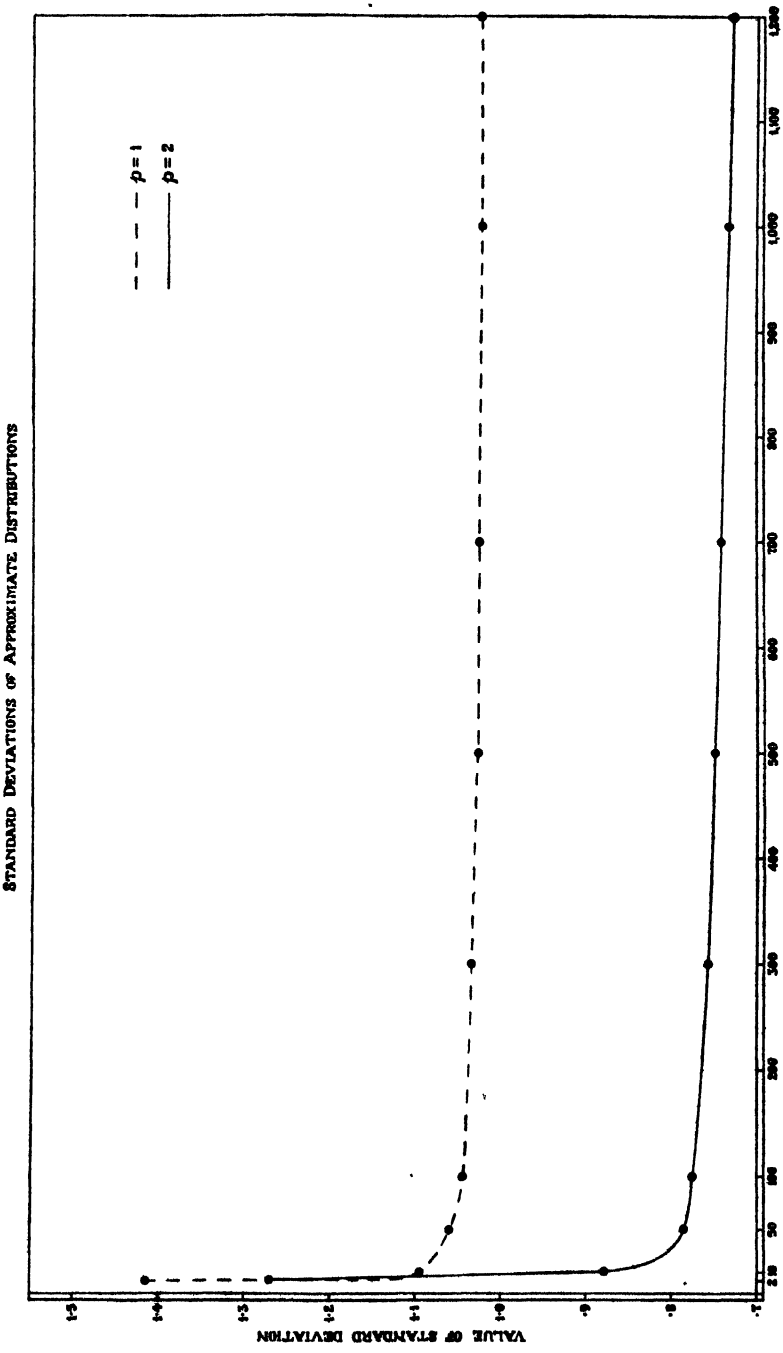

STANDARD DEVIATIONS OF APPROXIMATE DISTRIBUTIONS
— — — p = 1
——— p = 2
VALUE OF STANDARD DEVIATION
1·5
1·4
1·3
1·2
1·1
1·0
·9
·8
·7
SIZE OF SAMPLE
2
10
50
100
200
300
400
500
600
700
800
900
1,000
1,100
1,200

Diagram 3

CONSTANTS OF APPROXIMATE DISTRIBUTIONS. h AND y_0. $(p=1)$

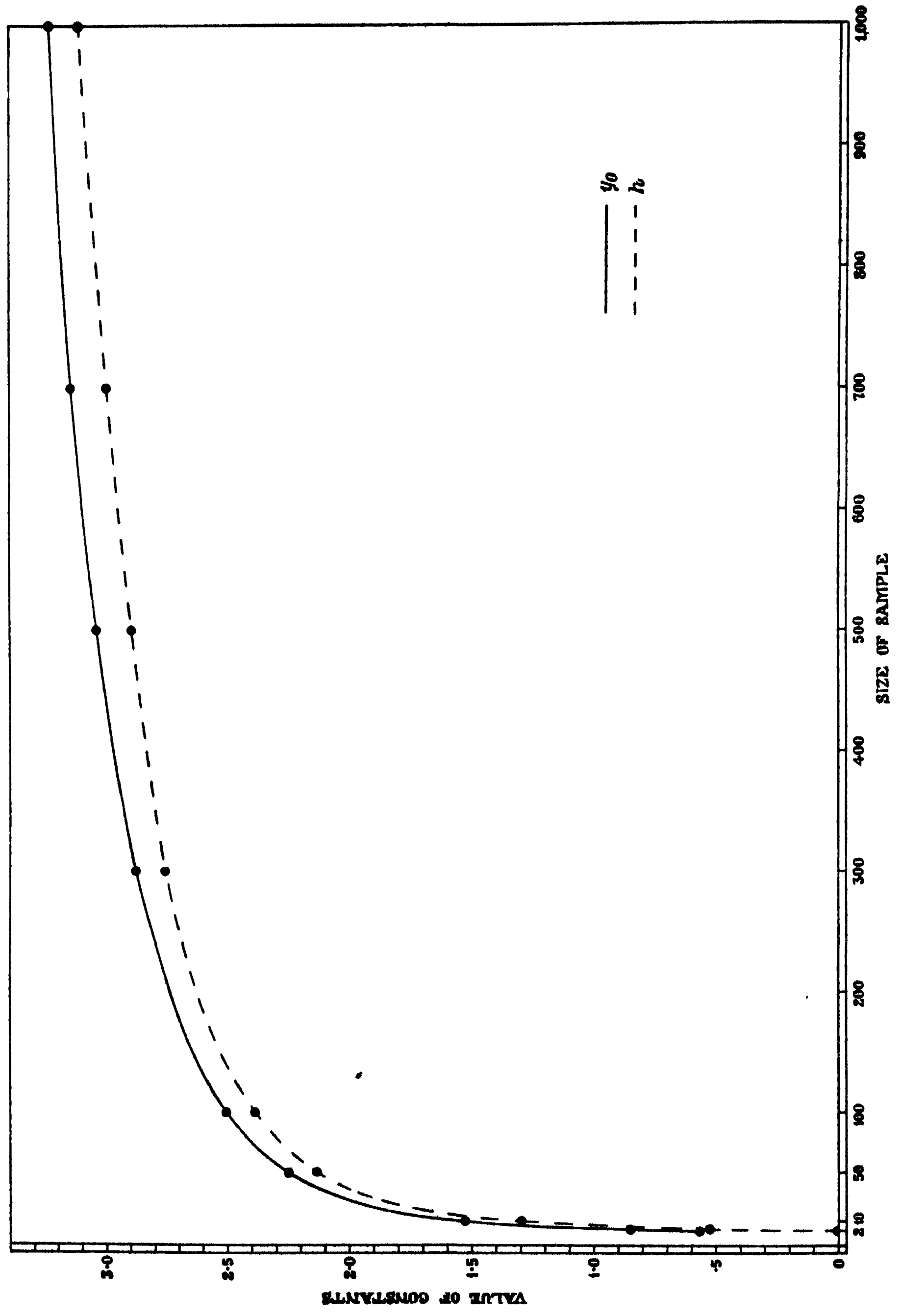


Diagram 4

CONSTANTS OF APPROXIMATE DISTRIBUTIONS. h AND y_0. $(p = 2)$

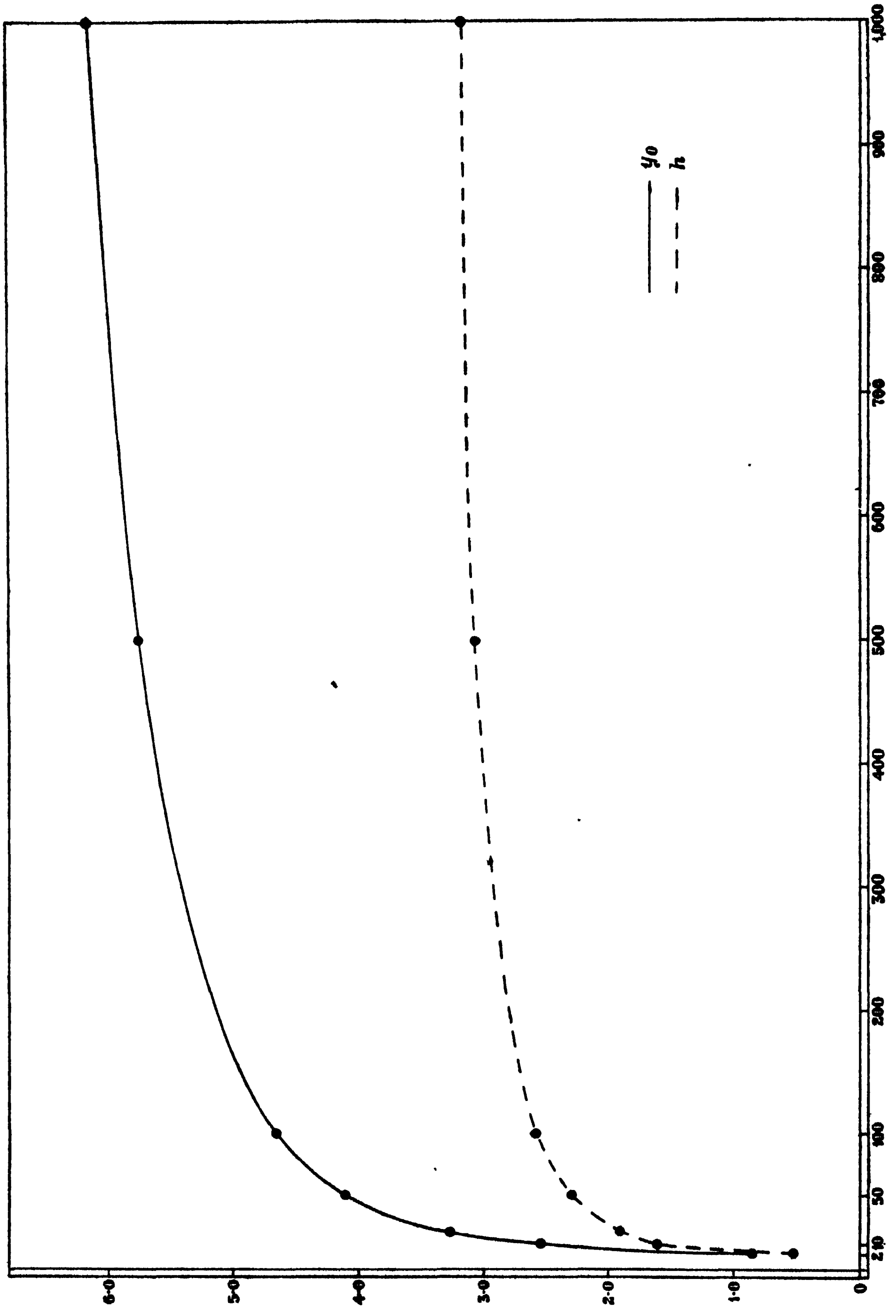


Diagram 5

Let $\chi_{2,1}$ be the mean of the differences, all taken positively, between pairs of individuals drawn at random from the original population, $\chi'_{2,1}$ the mean of the differences, all taken positively, between pairs of brothers drawn at random.

Then from the previous theory

$$\chi_{2,1} = \frac{2}{\sqrt{\pi}}\sigma \text{ ....(xxvii)}^*,$$

σ being the standard deviation of the original population and we have to find the mean value of $\chi'_{2,1}$.

Let the population of pairs of brothers be represented by the normal frequency surface

$$z = \frac{1}{2\pi\sigma^2\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2-2rxy+y^2}{\sigma^2}\right)}$$

and let us determine the mean value of $|x-y|$.

Let $$x = \frac{X-Y}{\sqrt{2}}, \quad y = \frac{X+Y}{\sqrt{2}},$$

then $$X = \frac{x+y}{\sqrt{2}}, \quad Y = \frac{x-y}{\sqrt{2}}$$

and $$|x-y| = |\sqrt{2}Y| = |Y'| \text{ say.}$$

The surface may then be written

$$\frac{1}{2\pi\sigma^2\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{(1-r^2)\sigma^2}[X^2+Y^2-r(X^2-Y^2)]} = \frac{1}{2\pi\sigma^2\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{X^2}{(1+r)\sigma^2}} e^{-\frac{1}{2}\frac{Y^2}{(1-r)\sigma^2}}.$$

The frequency distribution for Y may be obtained by integrating for X between $-\infty$ and ∞ namely:

$$\frac{1}{\sigma\sqrt{2\pi(1-r)}} e^{-\frac{1}{2}\frac{Y^2}{(\sigma\sqrt{1-r})^2}},$$

whence the frequency distribution for Y' is

$$z = \frac{1}{\sqrt{2\pi}\sigma\sqrt{2(1-r)}} e^{-\frac{1}{2}\frac{Y'^2}{[\sigma\sqrt{2(1-r)}]^2}} \text{ ....(xxviii).}$$

Therefore $$\chi'_{2,1} = \text{mean } |Y'| = \sqrt{\frac{2}{\pi}}\sigma\sqrt{2(1-r)} = \frac{2}{\sqrt{\pi}}\sigma\sqrt{1-r} \text{ ......(xxix)}^\dagger,$$

and $$\frac{\chi'_{2,1}}{\chi_{2,1}} = \sqrt{1-r} \text{ ....(xxx).}$$

---

[*] Cf. footnote, p. 101.

[†] This formula is a particular case for $\sigma_1 = \sigma_2$ of one first obtained by Pearson in the form

$$\frac{S_{x>y}\left\{\frac{x}{\sigma_1} - \frac{y}{\sigma_2}\right\}}{N} = \sqrt{\frac{1-r}{\pi}},$$

Drapers' Company Research Memoirs, No. IV. It can be used to determine r if σ_1, σ_2 are known, or are obtained from the observations. The method here proposed differs from Pearson's, by the fact that we eliminate the unknown standard deviation of the original population by taking the ratio of the mean of our two sets of differences.

It is worth while finding the probable error of r calculated from this formula. We shall obtain a first approximation only, assuming statistical differentials may be treated as mathematical differentials and that $\chi'_{2,1}, \chi_{2,1}$ are each obtained from averaging N pairs. Then

$$\left.\begin{aligned} \sigma^2\chi_{2,1} &= \frac{\sigma^2 d_2}{N} \\ \sigma^2\chi'_{2,1} &= \frac{\sigma^2 d'_{2,1}}{N} \end{aligned}\right\} \text{..............................} \quad \text{(xxxi)},$$

where $d_{2,1}$ denotes the difference, taken positively, between a pair drawn at random; $d'_{2,1}$ the difference, taken positively, between a pair of brothers.

Taking logarithmic differentials of (xxx)

$$-\frac{1}{2}\frac{\delta r}{(1-r)} = \frac{\delta\chi'_{2,1}}{\chi'_{2,1}} - \frac{\delta\chi_{2,1}}{\chi_{2,1}},$$

$$\frac{1}{4}\frac{(\delta r)^2}{(1-r)^2} = \frac{(\delta\chi'_{2,1})^2}{\chi'^2_{2,1}} + \frac{(\delta\chi_{2,1})^2}{\chi^2_{2,1}} - \frac{2\delta\chi_{2,1}\delta\chi'_{2,1}}{\chi_{2,1}\chi'_{2,1}}.$$

Summing for all possible samples and dividing by their number

$$\frac{1}{4}\frac{\sigma_r^2}{(1-r)^2} = \frac{\sigma^2\chi'_{2,1}}{\chi'^2_{2,1}} + \frac{\sigma^2\chi_{2,1}}{\chi^2_{2,1}} - \frac{2\,[\delta\chi_{2,1}\,\delta\chi'_{2,1}]}{\chi_{2,1}\chi'_{2,1}},$$

where the symbol [ ] denotes mean value.

But $$\sigma^2\chi_{2,1} = \frac{\sigma^2 d_{2,1}}{N} = \frac{1}{N}\left\{2\left(1-\frac{2}{\pi}\right)\right\}\sigma^2 \quad \text{..................(xxxii)*}$$

and $$[\delta\chi_{2,1}\delta\chi'_{2,1}] = 0,$$

because we are dealing with two independent processes of sampling.

It remains to find $\sigma^2\chi'_{2,1}$.

$$\sigma^2\chi'_{2,1} = \frac{\sigma^2 d'_{2,1}}{N} = \frac{1}{N}\sigma^2\,|Y'|$$

$$\frac{1}{N}\left\{2\int_0^\infty \frac{Y'^2 e^{-\frac{1}{2}\frac{Y'^2}{\{\sigma\sqrt{2(1-r)}\}^2}}\,dY'}{\sqrt{2\pi}\,\sigma\sqrt{2(1-r)}} - \chi'^2_{2,1}\right\}$$

$$= \frac{2}{N}\sigma^2(1-r) - \frac{1}{N}\frac{4}{\pi}\sigma^2(1-r)$$

$$\frac{2\sigma^2(1-r)}{N}\left\{1-\frac{2}{\pi}\right\} \quad \text{..................................(xxxiii)}.$$

Therefore $$\frac{\sigma^2\chi'_{2,1}}{\chi'^2_{2,1}} = \frac{\frac{1}{2}(\pi-2)}{N} = \frac{\sigma^2\chi_{2,1}}{\chi^2_{2,1}}$$

and $$\frac{\sigma_r^2}{(1-r)^2} = 4\left\{\frac{\sigma^2\chi'_{2,1}}{\chi'^2_{2,1}} + \frac{\sigma^2\chi_{2,1}}{\chi^2_{2,1}}\right\} = \frac{4(\pi-2)}{N},$$

or $$\sigma_r = 2(1-r)\sqrt{\frac{\pi-2}{N}} = 2{\cdot}137\,\frac{(1-r)}{\sqrt{N}} \quad \text{..............(xxxiv)},$$

---

* Cf. footnote, p. 111.

which gives the following ratios to the probable errors as they would be obtained by the product-moment method with the same value of N.

r	Ratio	r	Ratio
·1	1·943	·6	1·336
·2	1·781	·7	1·257
·3	1·644	·8	1·187
·4	1·526	·9	1·125
·5	1·425	1·0	1·068

Thus the probable error is always larger than that given by the product-moment method, but if a large enough number of pairs be taken there seems no reason why the method should not be practicable.

(3) In order that we might be able to obtain from the calculated r the probability of a deviation in excess or defect, we should really require to know not only the standard error but also the frequency distribution of r calculated in this way. This seems a very difficult problem. For large values of N it would presumably be approximately normal, but for smaller values it would seem to depend on the frequency distribution of differences. A knowledge of the frequency distribution of differences, therefore, may well be a first step in its solution.

At present our knowledge of the frequency distribution of differences is complete only for samples of two or three individuals, though good approximations have been obtained for larger samples. This together with consideration of probable errors suggests that it is for large numbers of small samples that Galton's Difference Problem is most likely to lead to practical results.

But it must be acknowledged, in conclusion, that the complicated theory of Galton's Difference Method makes it evident that very considerable difficulties will attend its application in practice, if an intelligible and well defined meaning is to be attached to the results obtained.

I am indebted to Professor Pearson for suggesting the subject of this investigation and for some valuable suggestions and criticism, and to Miss Ida M^cLearn for the preparation of the diagrams.

ON THE MULTIPLE CORRELATION OF BROTHERS, BEING A NOTE ON MR J. O. IRWIN'S MEMOIR, AND ON MY STATEMENT OF THE APPLICATION OF GALTON'S DIFFERENCE PROBLEM TO THE DETERMINATION OF THE DEGREE OF RELATIONSHIP OF BROTHERS, MADE IN AUGUST 1902. (*Biometrika*, Vol. I. p. 399.)

BY KARL PEARSON, F.R.S.

MR IRWIN'S paper does not seem to me to put the case for the application of Galton's Difference Problem to the subject of brothers adequately. I had not at the date his paper went to press examined my statement of 1902, and took on faith that what he wrote in his paper (ftn. p. 101) was in accordance with my remark of more than twenty years ago. Returning to that paper I find no ground for his interpretation of my statement. My words are the following:

> Anyhow I think we may say that if the individuals [i.e. a *pair*] be taken at random from a population, then the probable value* of the standard deviation of that population is nearly $\frac{23}{26}$ of the difference between the two individuals. Thus by averaging the differences between pairs of individuals taken at random we can obtain fairly readily an appreciation of the standard deviation, i.e. the variability of the general population. Further, if we take individuals not quite at random, but from correlated groups, e.g. pairs of brothers selected at random, the $\frac{23}{26}$th of the average difference of the pairs will be the standard deviation of the correlated groups, e.g. a group of brethren; hence the degree of relationship between such correlated individuals may be determined. This is only a suggestion of one of the many possible uses of Galton's difference problem. (*Biometrika*, Vol. I. p. 399.)

The paragraph may be somewhat obscurely worded, but I think the suggestion is definite in the fact that the pair has to be taken from a *group of brethren*, i.e. a group all of whom are offspring of the same parental pair. Mr Irwin supposes me to say that the two individuals are to be taken from a group all of whom have a brother of a given intensity of character, i.e. from the array of a fraternal correlation table. My statement says a "group of brethren" and thus interpreted there appears to me to be no error in it. A group of brethren differs from the general population in having all its ancestry right back the same. If σ be the standard deviation of the general population, then the standard deviation of such a group of brethren is $\sigma\sqrt{1-R^2}$, where R is the multiple correlation coefficient of the son on all his ancestors. If we pick out pairs of individuals at random from a group of brothers, and from the general population, then the ratio of the mean differences of such respective pairs is such that:

$$\frac{\text{Mean difference of fraternal pairs}}{\text{Mean difference of general population pairs}} = 1 - R^2.$$

[* The use of the term "probable value" in the light of our present knowledge is incorrect. The statement as to averaging differences is correct. But the mean or average value is not the probable value, for the distribution is known to be skew, as is well illustrated in Mr Irwin's paper.]

It is thus possible to find R by this process, and R involves at once a determination of r the so-called "correlation of brothers" as found by a fraternal correlation table.

The difficulty of the process in the case of *human* families is their small size; this renders it impossible to draw pairs from an adequately large group of brothers, who would have $\sigma\sqrt{1-R^2}$ for their standard deviation. But this difficulty (which does not exist in the case of some other species) can be got over as follows:

Suppose we take samples of size n from a population of standard deviation Σ following the normal law of distribution. Then the distribution curve of the standard deviation s in the samples is known to be

$$z = \text{const.} \times \frac{s^{n-2}}{\Sigma^{n-1}} e^{-\frac{ns^2}{2\Sigma^2}},$$

and the chance of a standard deviation falling between s and $s + ds$ is proportional to

$$\frac{s^{n-2}}{\Sigma^{n-1}} e^{-\frac{ns^2}{2\Sigma^2}} ds.$$

But this must also be the chance of the square of the standard deviation s^2 or μ_2 having a value between s^2 and $(s + \delta s)^2$, or between μ_2 and $\mu_2 + \delta\mu_2$. Expressing the result in terms of μ_2, we have the chance proportional to

$$\left(\frac{\mu_2}{\Sigma^2}\right)^{\frac{n-3}{2}} e^{-\frac{n\mu_2}{2\Sigma^2}} \frac{d\mu_2}{\Sigma^2},$$

or the frequency curve for the distribution of μ_2 is

$$z_y = \frac{z_0}{\Sigma^2}\left(\frac{\mu_2}{\Sigma^2}\right)^{\frac{n-3}{2}} e^{-\frac{n\mu_2}{2\Sigma^2}}.$$

The mean value of μ_2 in samples, or $\bar{\mu}_2$, is accordingly

$$\bar{\mu}_2 = \Sigma^2 \frac{\displaystyle\int_0^\infty \left(\frac{\mu_2}{\Sigma^2}\right)^{\frac{n-3}{2}} \frac{\mu_2}{\Sigma^2} e^{-\frac{n\mu_2}{2\Sigma^2}} \frac{d\mu_2}{\Sigma^2}}{\displaystyle\int_0^\infty \left(\frac{\mu_2}{\Sigma^2}\right)^{\frac{n-3}{2}} e^{-\frac{n\mu_2}{2\Sigma^2}} \frac{d\mu_2}{\Sigma^2}}$$

$$= \frac{2\Sigma^2}{n} \frac{\displaystyle\int_0^\infty z^{\frac{1}{2}(n-1)} e^{-z} dz}{\displaystyle\int_0^\infty z^{\frac{1}{2}(n-3)} e^{-z} dz} = \frac{2\Sigma^2}{n} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

$$\frac{n-1}{\;} \Sigma^2$$

Hence the mean standard deviation squared of *pairs* of brothers is obtained by taking $n = 2$, or is equal to $\frac{1}{2}\Sigma^2 = \frac{1}{2}\sigma^2(1 - R^2)$.

Now let x_1 and x_2 be two individuals, then their mean $=\frac{1}{2}(x_1+x_2)$ and their standard deviation squared equals

$$\frac{1}{2}\{((x_1-\tfrac{1}{2}(x_1+x_2))+(x_2-\tfrac{1}{2}(x_1+x_2))^2\}$$
$$=\frac{1}{4}(x_1-x_2)^2.$$

Accordingly we want the mean value of $\frac{1}{4}(x_1-x_2)^2$ for all pairs, say N, and this equals

$$\frac{\frac{1}{4}S\{x_1-\bar{x}-(x_2-\bar{x})\}^2}{N}$$
$$=\frac{1}{4}(\sigma_1^2+\sigma_2^2-2r\sigma_1\sigma_2),$$

where r is the so-called correlation of brothers and σ_1 is the standard deviation of the first, σ_2 of the second brother. Taking these equal, we have both equal σ; and the mean standard deviation squared of pairs of brothers $=\frac{1}{2}\sigma^2(1-r)=\frac{1}{2}\sigma^2(1-R^2)$. Hence $r=R^2$, or the correlation of brothers is equal to the squared multiple correlation coefficient of the offspring on all ancestry. A knowledge of R gives r; and Mr Irwin translates my $1-R^2$ into $1-r^2$ and supposes I meant pairs from an array and not pairs from a group of brothers as I wrote.

We may now examine the matter from another aspect. Consider the variate x_0 in correlation with the variates $x_1, x_2, \ldots x_n$, the correlation coefficients being given by r_{uv} between the uth and vth variates. Then it is well known that the multiple correlation coefficient of x_0 on $x_1, x_2, \ldots x_n$ is given by R_n, where

$$R_n^2=1-\frac{\Delta}{\Delta_{00}},$$

Δ being the determinant

$$\Delta=\left|\begin{array}{l} 1\ ,\ r_{01}\ ,\ r_{02}\ ,\ldots r_{0n} \\ r_{10}\ ,\ 1\ ,\ r_{12}\ ,\ldots r_{1n} \\ r_{20}\ ,\ r_{21}\ ,\ 1\ ,\ldots r_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{n0}\ ,\ r_{n1}\ ,\ r_{n2}\ ,\ldots 1 \end{array}\right|$$

and Δ_{00} the minor of the leading constituent. If Δ takes the form

$$\begin{array}{l} 1,\ r,\ r,\ldots r \\ r,\ 1,\ \rho,\ldots \rho \\ r,\ \rho,\ 1,\ldots \rho \\ \ldots\ldots\ldots\ldots\ldots \\ r,\ \rho,\ \rho,\ldots 1 \end{array}$$

it is easy to show that

$$R_n=r\sqrt{\frac{n}{1+(n-1)\rho}}.$$

Accordingly if $\rho=r$ and n be made indefinitely great, we have

$$R_\infty=r\sqrt{\frac{1}{\rho}}=\sqrt{r}.$$

The form thus taken by R is, however, the form suitable to an indefinitely great group of brothers. We learn accordingly that:

(*a*) The multiple correlation coefficient of the offspring on an indefinitely extended ancestry is equal to the multiple correlation coefficient of one brother on an indefinitely great number of brothers.

(*b*) Either of these multiple correlation coefficients is the square root of the ordinary correlation coefficient for two brothers.

(*c*) The simple correlation coefficient for two brothers is numerically of about the order 0·5; it follows that the multiple fraternal correlation is of an order somewhat over 0·7. It is this multiple fraternal correlation coefficient which is directly determined, when we take the ratio of the mean difference between pairs of brothers to the mean difference between pairs of individuals from the general population.

Further, the importance of the multiple fraternal correlation (R), if determined, is very great, for on the assumption of the Law of Ancestral Heredity—i.e. the decadence of individual ancestral correlations in geometric progression—it enables us to write down with comparative ease the expression for the mean value of the offspring with a given ancestry. The fundamental formulae for the Law of Ancestral Heredity are given in *Biometrika*, Vol. VIII. pp. 239—240, and may be briefly recapitulated here. x_0 is the character value in the individual offspring in the last generation. X_p is the character of the "mid-parent" in the pth ancestral generation, $\bar{x}_0$ and $\bar{X}_p$ are the corresponding means and σ_0, Σ_p the corresponding standard deviations; σ_p the standard deviation of individuals in the pth ancestral generation; ρ_{p-q} the *average* correlation of individual ancestors in the pth generation with their offspring in the qth generation ($p > q$); ϵ_p the coefficient of assortative mating in the parents of the pth ancestral generation, ρ_{p-q} the *average* coefficient of correlation between individual ancestors of the pth and individual offspring of the qth ancestral generation, and r_{p-q} the correlation between the mid-parent of the pth and the mid-parent of the qth generation. Then the multiple regression formula of the offspring on the series of ancestral mid-parents being

$$x_0 - \bar{x}_0 = S\left\{c_p \frac{\sigma_0}{\Sigma_p}(X_p - \bar{X}_p)\right\},$$

with a variability

$$\Sigma_0 = \sigma_0 \sqrt{1 - S(c_p r_{p0})},$$

and the geometrical decrease of ρ_{p-q} being assumed, it is shown that

$$\Sigma_p = \frac{\sigma_p}{(\sqrt{2})^p}\sqrt{1 + \epsilon_p},$$

$$r_{p-q} = (\sqrt{2})^{p-q}\rho_{p-q}/\sqrt{(1+\epsilon_p)(1+\epsilon_q)}.$$

Accordingly if assortative mating be constant throughout the generations, r_{p-q} follows the geometric law if ρ_{p-q} does. Then taking

$$r_{p-q} = \alpha\beta^{p-q}, \quad c_p = \gamma\eta^p,$$

we find

$$\beta - \eta = \gamma\eta,$$

$$\eta^2 - \eta\,\frac{(1 - 2\alpha\beta^2 + \beta^2)}{\beta(1-\alpha)} + 1 = 0,$$

which give γ and η in terms of α and β. Further,

$$\Sigma_0 = \sigma_0\sqrt{1 - \frac{\gamma\eta\alpha\beta}{1 - \eta\beta}} = \sigma_0\sqrt{1 - R^2},$$

or

$$R^2 = \frac{\gamma\eta\alpha\beta}{1 - \eta\beta} = r,$$

the fraternal correlation as we have shown on p. 131 above.

Accordingly a knowledge of R^2 and the equations for η and γ above enable us to find β, γ and η in terms of α or $\alpha\beta$. But for parental correlation we have

$$\alpha\beta = r_1 = \sqrt{2}\,\rho_1/(1 + \epsilon),$$

or $\alpha\beta$ is known in terms of the mean (ρ_1) of the two parental correlations and the coefficient ϵ of assortative mating.

We will write $$\rho_1/(1 + \epsilon) = \rho_1'.$$

Substituting for $\gamma\eta$ in the value for R^2, we have

$$R^2 = \frac{(\beta - \eta)\,\alpha\beta}{1 - \eta\beta},$$

whence we find

$$\eta = \frac{R^2 - \alpha\beta^2}{\beta(R^2 - \alpha)} = \frac{R^2 - \beta\sqrt{2}\,\rho_1'}{\beta R^2 - \sqrt{2}\,\rho_1'}.$$

Substituting this value of η in the quadratic for η above, we find

$$\beta^2 - 2\beta\sqrt{2}\,\rho_1' + \frac{2\rho_1'^2 - R^2(1 - R^2)}{R^2} = 0,$$

or

$$\beta = \sqrt{2}\,\rho_1' \pm \sqrt{\frac{(1 - R^2)(R^2 - 2\rho_1'^2)}{R^2}}.$$

We shall consider the alternate roots later.

Then $$\alpha = \sqrt{2}\,\rho_1'/\beta,$$

and

$$\rho_{p-q} = \rho_1\left(\frac{\beta}{\sqrt{2}}\right)^{p-q-1}$$

gives the geometric law of degradation of the average individual ancestral correlations. Finally

$$\gamma\eta = \frac{(\beta^2 - 1)R^2}{\beta R^2 - \sqrt{2}\,\rho_1'}, \text{ and } \eta = \frac{R^2 - \beta\sqrt{2}\,\rho_1'}{\beta R^2 - \sqrt{2}\,\rho_1'}$$

enable us to determine

$$c_p = \gamma\eta \cdot \eta^{p-1}$$

and complete the solution for the multiple regression equation of offspring on ancestry.

Thus, a knowledge of:

(*a*) Parental correlations;

(*b*) Fraternal correlations;

(*c*) Coefficient of assortative mating, and the assumptions

(i) that the population is stable

$$\sigma_p = \sigma_0, \text{ for all values of } p,$$
$$\epsilon_p = \epsilon \quad \text{,,} \quad \text{,,}$$

(ii) that the individual ancestral correlations follow a geometric law of degradation

enable us to determine the multiple regression equation, without determining the individual ancestral correlations. There is, however, an exception which is so important that it needs special consideration. Our equations become indeterminate when $\beta = 1$. We will deal with this case.

Case of equality of all coefficients of correlation. When $\beta = 1$, $r_{p-q} = \alpha$, whatever the values of p and q. All the mid-parents have the same correlations with each other and with any generation of offspring. We therefore need the multiple regression equation, when the fundamental determinant is

$$\Delta = \begin{vmatrix} 1, & r, & r, \ldots r \\ r, & 1, & r, \ldots r \\ r, & r, & 1, \ldots r \\ & & \\ r, & r, & r, \ldots 1 \end{vmatrix} \quad \text{to } n+1 \text{ rows and columns.}$$

Add up the last n rows, divide their sums by $\dfrac{1 + \overline{n-1}\, r}{r}$ and subtract from the first row. We have

$$\Delta = \begin{vmatrix} 1 - \dfrac{nr^2}{1 + \overline{n-1}\, r}, & 0, & 0, \ldots 0 \\ -r & , 1, & r, \ldots r \\ r & , r, & 1, \ldots r \\ \ldots\ldots\ldots\ldots & \ldots\ldots & \ldots\ldots\ldots \\ r & , r, & r, \ldots 1 \end{vmatrix}$$

or
$$\Delta = \left(1 - \frac{nr^2}{1 + (n-1)r}\right), \ \Delta_{00} = \frac{(1-r)(1+nr)}{1+(n-1)r}\Delta_{00}.$$

Similarly
$$\Delta_{00} = \frac{(1-r)(1+(n-1)r)}{1+(n-2)r}\Delta_{0011},$$

where Δ_{0011} is the minor obtained by cancelling the first two rows and first two columns.

Now let us endeavour to find

$$\Delta_{01} = - \begin{matrix} r, & r, & r, \ldots r \\ r, & 1, & r, \ldots r \\ r, & r, & 1, \ldots r \\ \ldots\ldots & \ldots\ldots & \ldots\ldots \\ r, & r, & r, \ldots 1 \end{matrix}$$

Add up the last $n-1$ rows, divide their sums by $\frac{1+(n-2)r}{r}$ and subtract from the first row. We have

$$\Delta_{01} = -\begin{vmatrix} r - \frac{(n-1)r^2}{1+(n-2)r}, & 0, & 0, \ldots & 0 \\ r \quad , & 1, & r, \ldots & r \\ r \quad , & r, & 1, \ldots & r \\ \\ r \quad , & r, & r, \ldots & 1 \end{vmatrix}$$

$$= -\left(r - \frac{(n-1)r^2}{1+(n-2)r}\right)\Delta_{0011}$$

$$= -\frac{r(1-r)}{1+(n-2)r}\Delta_{0011}.$$

Substituting the value already found for Δ_{0011}, we have

$$\frac{\Delta_{01}}{\Delta_{00}} = -\frac{r}{1+(n-1)r} = \frac{\Delta_{0p}}{\Delta_{00}}$$

by the symmetry of the determinant.

Accordingly the multiple regression equation for the mean value of an array of x_0's, when determined from a series of variates $x_1, x_2, \ldots x_n$ of standard deviations $\sigma_1, \sigma_2, \ldots \sigma_n$, when all the correlations are the same, is

$$\tilde{x}_0 - \bar{x}_0 = \sigma_0 \overset{n}{\underset{1}{S}} \frac{r}{1+(n-1)r} \frac{x_p - \bar{x}_p}{\sigma_p},$$

where $\Sigma_0 = \sigma_0\sqrt{1-R^2}$ gives the variability, and $R^2 = \frac{nr^2}{1+(n-1)r}$ is the multiple correlation coefficient *.

When n becomes indefinitely large $R^2 = r$ and the regression coefficient becomes the sum of an infinitely large series of infinitely small terms. That this series will take a finite limit is clear; for example, if we take $(x_p - \bar{x}_p)/\sigma_p = \text{constant} = q$, then $\tilde{x}_0 - x_0 = \sigma_0 \frac{rnq}{1+(n-1)r} = q\sigma_0$ in the limit.

Returning to the case of predicting the probable value of a character from a knowledge of the ancestral values, we have $r = r_{p-q} = \alpha\beta^{p-q} = \alpha = \sqrt{2}\,\rho_1'$. Thus

$$\tilde{x} - \bar{x}_0 = \sigma_0 \frac{\sqrt{2}\,\rho_1'}{1+(n-1)\sqrt{2}\,\rho'} \overset{n}{\underset{1}{S}} \frac{X_p - \bar{X}_p}{\Sigma_p},$$

$$\Sigma_0 = \sigma_0\sqrt{1-R^2}, \text{ where } R^2 = \frac{2n\rho_1'^2}{1+\sqrt{2}(n-1)\rho_1'}.$$

The previous remark holds good here, that as n becomes very large we have an indefinitely large number of indefinitely small terms. But there is no advantage in taking n indefinitely large. When $n = 10$, excluding cousin marriages, there will be 2048 ancestors in the 10th ancestral generation and the "mid-parent" provided by them will be essentially equal to the mean of that generation or $X_p = \bar{X}_p$.

---

* I do not remember publishing this result before, but have given it occasionally in lecture to illustrate multiple correlation.

We may note that in this case

$$\rho_{p-q} = \rho_1 \left(\frac{\beta}{\sqrt{2}}\right)^{p-q-1} = \rho_1 \left(\frac{1}{\sqrt{2}}\right)^{p-q-1},$$

or the radix of the geometrical progression is ·707107 [*].

If we take our system of equal correlations, the deduction of the probable character of one brother from n others, then $\sigma_0 = \sigma_p$ for all values of p, and if r be the simple fraternal correlation

$$\tilde{x}_0 - \bar{x}_0 = \frac{nr}{1 + (n-1)r}\left(\frac{S(x_p)}{n} - \bar{x}_0\right),$$

$$\Sigma_0 = \sigma_0 \sqrt{1 - R'^2},$$

where

$$R'^2 = \frac{nr^2}{1 + (n-1)r}.$$

When n becomes infinite we have the array of brothers or sisters, which *could* form a family. But this must be the array of offspring from the same total ancestry, i.e. for $n = \infty$,

$$R'^2 = R^2,$$

or

$$r = \sqrt{2}\,\rho_1'.$$

Here $\rho_1' = \rho_1/(1+\epsilon)$.

Taking stature, forearm and span, we find the mean value of ϵ is ·235667 [†]. The mean parental inheritance for the same data was ·45925 [‡]. Hence $\rho_1' = $ ·371,6616 and $\sqrt{2}\,\rho_1' = $ ·5256. The mean value of r the correlation coefficient in the corresponding tables for stature, forearm and span in pairs of brothers and in pairs of sisters was ·5250 [§], which is sensibly in accord with the above result. The advantage of such a formula as that just given is that it enables us to deduce fraternal from parental and assortative mating correlations, or *vice versa*. The formula, however, essentially depends on β being nearly equal to unity, a result which would flow from the hypothesis that the associations of the mid-parents of each generation of ancestry to each other and to the offspring are of the same intensity. I now propose to develop the formulae provided in this note numerically from the data given by my family measurements of 1893. The cards then issued were limited to four adult siblings in each family, and so in many cases it was only possible to measure one brother and one sister. Thus the pairs of brothers and the pairs of sisters were not as numerous as we could have wished, while we had ample material for parental inheritance and for brother-sister pairs. To illustrate points in this present note and in Mr Irwin's preceding paper, I have confined myself to the single character of stature in pairs of brothers and pairs of sisters. Table I gives the differences in stature of 341 pairs of brothers and 482 pairs of sisters arranged in $\frac{1}{4}''$ units[*]. I first took the differences given in this table, squared them and took their weighted mean. This had to be divided by 64, because the units are in quarter inches and because what we need (see p. 131) is the mean value of $\frac{1}{4}(x_1 - x_2)^2$. In this way I found

---

[*] For example the mean parental correlation coefficient being taken at ·460 (*Biometrika*, Vol. II. p. 378) we have ·3253, ·2300, ·1626, ·1150, ·0813, etc. for the successive grandparental average correlations. These are by no means out of accord with our somewhat limited experience.

[†] *Biometrika*, Vol. II. p. 378.

[‡] *Ibid.* p. 378.

[§] *Ibid.* p. 387.

$$\tfrac{1}{2}\sigma^2 (1 - R^2) = 1\cdot557{,}345, \text{ for brothers,}$$

$$= 1\cdot597{,}267, \text{ for sisters.}$$

But σ^2 in inches was 6·85866 for 682 brothers and 6·94701 for 964 sisters.

Calculating R^2, from the above values, I found

For Brothers: $R^2 = \cdot545{,}875$; for Sisters: $R^2 = 540{,}157$.

Accordingly, $R = \cdot7388$ for brothers and $= \cdot7350$ for sisters; these are the values of the multiple correlation coefficients for an indefinitely large number of brothers and sisters respectively.

We can approach the subject from another side. The correlation coefficient for the 682 pairs of brothers was ·5196 and for the 964 pairs of sisters[†] was ·5370. Hence $R = \sqrt{r}$ gives us $R = \cdot7208$ and $\cdot7328$ for brothers and sisters respectively. These results are not bad, better in the case of sisters than of brothers; we must bear in mind, however, that our value for the mean squared standard deviation of pairs depends on selecting at random from a *normal* distribution.

TABLE I.

Distribution of Differences of Stature between Pairs of Brothers and Pairs of Sisters.

Frequency of Pairs.

Differences in quarter inches	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
Brother-Pairs	11·5	24	29·5	17	35·5	18·5	24·5	22·5	31	15	18	10·5	18·5	5
Sister-Pairs	28	35	37	24	45	26	32	25	35·5	16	39	14·5	25·5	9
Differences in quarter inches	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*
Brother-Pairs	10	5	11	6·5	5·5	5	3·5	2·5	3	—	3	2	1	—
Sister-Pairs	18·5	15	17	8	6	4	2	2	5	2	3	4	—	—
Differences in quarter inches	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*	*39*	*40*	Totals
Brother-Pairs	1	—	—	—	—	—	1	—	—	—	—	—	1	341
Sister-Pairs	—	1	—	1	1	—	—	—	—	—	1	—	—	482

---

[*] I have to thank my colleague, Miss E. M. Elderton, for taking out this table from the original data cards and for finding several constants for me.

[†] The tables were rendered symmetrical, as there was no reason to distinguish for our present purpose between younger and elder siblings, so that they contained 682 and 964 pairs respectively.

We now proceeded to take the mean difference of pairs. This was found to be $2''\cdot 03391$ for pairs of brothers and $2''\cdot 01686$ for pairs of sisters, the two standard deviations being $2''\cdot 61890$ and $2''\cdot 63572$ respectively.

Calculating $r\,(=R^2)$ from the formula*

$$\left(\frac{\text{Mean difference of pairs}}{\text{S.D. of population}}\right)^2 = \frac{4}{\pi}(1-r),$$

we have

$$r = \cdot 526{,}289 \text{ for brothers, } = \cdot 540{,}119 \text{ for sisters,}$$

the correlation table values being

$$r = \cdot 5196 \text{ for brothers, } = \cdot 5370 \text{ for sisters.}$$

The former values give

$$R = \cdot 7255 \text{ for brothers and } \cdot 7349 \text{ for sisters.}$$

Thus we have the following values of R the multiple fraternal correlation coefficient:

	Pairs of Brothers	Pairs of Sisters
From mean square S.D. of pairs ...	·7388	·7380
From correlation table value of r ...	·7208	·7328
From mean difference of brothers ...	·7255	·7349
Mean of three methods	·7284	·7342
Mean of results for brothers and sisters	·7311	
Corresponding to an average individual fraternal correlation	·5345	

It will be clear from this numerical investigation that the value of the multiple correlation coefficient can be found with reasonable accuracy by taking the squares of the differences of brothers and of non-brothers, and that

$$\frac{\text{Mean square difference of character in brothers}}{\text{Mean square difference of character in non-brothers}} = 1 - R^2.$$

Or, again we may use:

$$\frac{\text{Mean difference of character in brothers}}{\text{Mean difference of character in non-brothers}} = \sqrt{1-R^2}.$$

The simple correlation coefficient may be found from the multiple correlation coefficient by the relation $r = R^2$.

We may now proceed to determine the constants of the multiple regression formulae on the basis of the values of the multiple correlation coefficients determined from the mean square standard deviations of pairs, i.e.

$$R^2 = \cdot 545{,}875 \text{ for brothers, } = \cdot 540{,}157 \text{ for sisters.}$$

We have first, however, to determine ρ_1 and ρ_1' for stature. Taking the mean of the two parental correlations for stature, these give

$$\rho_1 = \cdot 5040 \text{ for sons, and } = \cdot 5085 \text{ for daughters.}$$

---

* "On Further Methods of measuring Correlation." *Biometric Laboratory Publications*, IV. p. 5.

Now the coefficient of assortative mating in stature for this same material is $\epsilon = \cdot 2804$. Accordingly

$\rho_1' = \rho_1/(1+\epsilon) = \cdot 393{,}627$ for sons, and $= \cdot 397{,}142$ for daughters,
$\sqrt{2}\,\rho_1' = \cdot 556{,}673$ for sons, and $\cdot 561{,}644$ for daughters,
$2\rho_1'^2 = \cdot 309{,}885$ „ „ $\cdot 315{,}444$ „ „

From these values we find from the quadratic (p. 133):

For Sons	For Daughters
$\beta = \cdot 999{,}759$,	$\beta = \cdot 999{,}024$,
or, $= \cdot 113{,}587$,	or, $= \cdot 124{,}264$.

Now the individual ancestral correlations will be given by

$$\rho_p = \cdot 5040 \left(\frac{\beta}{\sqrt{2}}\right)^{p-1}, \quad \rho_p = \cdot 5085 \left(\frac{\beta}{\sqrt{2}}\right)^{p-1}, \text{ respectively.}$$

Or, by

$$\rho_p = \cdot 5040\,(\cdot 706{,}936)^{p-1}, \quad \rho_p = \cdot 5085\,(\cdot 706{,}417)^{p-1},$$

or,

$$\rho_p = \cdot 5040\,(\cdot 080{,}318)^{p-1}, \quad \rho_p = \cdot 5085\,(\cdot 087{,}868)^{p-1}.$$

The latter values for the individual ancestral correlations show results with a ratio of degradation far too small for our experience of grandparental and great-grandparental correlations. The second roots of the quadratic—those from the minus sign—must be discarded. Or, we conclude that

$\beta = \cdot 999{,}759$ for sons, and $= \cdot 999{,}024$ for daughters.

But values obtained from statistical data and differing only so slightly from unity suggest most cogently that the true value of β is *unity*, or that actually our multiple regression falls into the special case dealt with on pp. 134—135. The biological significance of this result—applied to *somatic* character measurements—is very considerable. In statistical language we may state the principle as follows: The correlation of any "mid-parent" with every other mid-parent in any ancestral generation and with the offspring is the same. Symbolically $\rho_{p-q} = \alpha$, and is independent of p and q.

Now $\alpha = \sqrt{2}\,\rho_1/\{\beta(1+\epsilon)\} = \sqrt{2}\,\rho_1' = \cdot 5567$ for sons and $\cdot 5616$ for daughters.

Had there been no assortative mating ($\epsilon = 0$) we should have found

$\alpha = \cdot 7128$ for sons and $\cdot 7193$ for daughters.

Hence according to the intensity of assortative mating there may be considerable variation in the value of this constant mid-parental correlation.

The values obtained for mean individual ancestral correlations are as follows—of course for stature:

Offspring and	Sons	Daughters
Parent	·5040	·5085
Grandparent	·3563	·3592
Great-Grandparent	·2519	·2538
Great-great-Grandparent	·1781	·1793
Great-great-great-Grandparent	·1259	·1266
Great-great-great-great-Grandparent ...	·0890	·0895

We have no ancestral correlations for any accurately measurable character in man. Unfortunately in the case of the coat-colour in horses and eye-colour in man in the higher relationships lack of data did not permit of a separation of the sexes. The most we can do on comparison is to say that values more or less similar to the above have been found, but that until we have measurable characters in half-a-dozen generations, we can only assert that the above values for stature do not present any great improbability as to the rate of degradation of ancestral influence.

Turning to the fundamental multiple regression equation for $\beta = 1\cdot$ on p. 135, we note that for n ancestral generations:

$$\tilde{x}_0 - \bar{x}_0 = \sigma_0 \frac{\cdot 5567}{1 + (n-1)\cdot 5567} \overset{n}{\underset{1}{S}} \frac{X_p - \bar{X}_p}{\Sigma_p} \text{ for sons,}$$

and

$$\tilde{x}_0 - \bar{x}_0 = \sigma_0' \frac{\cdot 5616}{1 + (n-1)\cdot 5616} \overset{n}{\underset{1}{S}} \frac{X_p - \bar{X}_p}{\Sigma_p} \text{ for daughters.}$$

Here

$$\Sigma_p = \sigma_p \left(\frac{1}{\sqrt{2}}\right)^p \sqrt{1 + \epsilon}$$

$$= 1\cdot 13155\,(\cdot 70711)^p\,\sigma_p$$

is the standard deviation of the pth mid-parents and σ_p is the standard deviation of male individuals in the pth ancestral generation. All female individuals are to be reduced to their male equivalents by multiplying their deviation from the female mean of their generation by the factor σ_p/σ_p', where σ_p' is the female standard deviation. $\bar{X}_p'$ is the mean of all possible mid-parents in the pth generation thus reduced. This definition of mid-parent and mid-parental mean is not an arbitrary one, but flows at once from the general principles of linear regression. For our particular data σ_0 is $2''\cdot 61890$ for sons and σ_0' is $2''\cdot 63572$ for daughters. While, as before stated, the substitution of Σ_p would appear formally to give a divergent series, this is not so actually. Every $X_p - \bar{X}_p$ must be estimated in terms of its special standard deviation, and while Σ_p gets smaller and smaller so does $X_p - \bar{X}_p$. This flows from the simple fact that without special inbreeding it is impossible to obtain several hundred ancestors in the 7th or 8th generation all deviating by the same considerable quantity from the mean of their generation. It is easy to have two exceptional parents, but not easy to have a multitude of exceptional ancestors of a high ancestral order. For particular purposes it will always be adequate to take n a simple digit in the above formulae.

To some biologists this *Note*—if they consent to read it—may appear a mere juggling with figures, which no doubt we shall be informed are valueless as data because, it will be asserted, they were not collected by a biologist! But at bottom there is a broad biological principle in the matter dealt with, which is really not beyond the comprehension of the non-mathematical. Let us take all the possible offspring of a human pair as a single group G_0, the parents as a group G_1, the grandparents as a group G_n and the ancestors in the pth generation as a group G_p. Then a study of the reduced variability of group G_0 as compared with the general population of their generation in the case of stature has shown us that if we treat the groups G_0, G_1, ... G_n,... as individuals, these individuals are all equally associated with the ultimate offspring G_0 or with any other of these ancestral groups. It is precisely as if each group of co-generational ancestry including the offspring formed a single individual, and these single individuals like a group of brothers had the same degree of resemblance to each other. Is the individual more like his parents than his grandparents, or more like his grandparents than his great-grandparents? He is more like his *individual* parent than his *individual* grandparent, and more like the latter than his *individual* great-grandparent and so on (see p. 139). But if we drop the word *individual*, and replace it by the group as a whole, then the offspring resemble in an equal degree each successive ancestral group as represented by the appropriate "mid-parent." The individual offspring is as like to his grandparents, or indeed to his 2048 group of ancestors in the 10th preceding generation, as he is to his direct parental pair. Such is the conclusion we can draw, for at least one somatic character, by a full investigation of the multiple correlation coefficient of an indefinitely large group of brothers.

I have endeavoured in this *Note* on Mr Irwin's paper to express the ideas of which I was more or less conscious when I wrote my paragraph on the application of Galton's Difference Problem to pairs of brothers in the summer of 1902. My wording was possibly obscure, but I was convinced that the suggestion contained more than Mr Irwin is now inclined to read into it, and I am very grateful to him for bringing me back to the study of my note on the subject in the first volume of this Journal. It was not till I had read it carefully again, that my ideas of 1902 came back to me more clearly, and he has enabled me at last to put on record what I meant then by the paragraph which I venture to think he has misread.

HEIGHTS AND WEIGHTS OF PATIENTS IN MENTAL HOSPITALS.

By ROSA GREENWOOD, CECILY M. THOMPSON and HILDA M. WOODS.

INTRODUCTION.

IN 1923 a Committee appointed by the Chairman of the Board of Control to consider the diets of County and Borough Mental Hospitals requested a number of Medical Superintendents to furnish particulars of heights, of weights on admission and subsequently of ages of samples of their patients.

The primary object of this request was to ascertain whether any relation could be discerned between the weights of adult patients and the computed energy values of the standard dietaries of the institutions sampled. For reasons stated in the Report of the Committee this object was not satisfactorily attained, but, as the returns provided a larger mass of statistics of heights and weights than has so far been published in reports or papers upon persons treated in English Mental Hospitals, it was thought expedient to analyse the data as completely as practicable. A note on the chief results was printed among the appendices to the Report of the Committee on Diets. That note was written by Dr Major Greenwood, a member of the Committee, under whose general supervision this analysis was made.

The inquiry has been conducted with the approval of the Committee on Quantitative Factors of Human Nutrition (Medical Research Council), and the expenses have been defrayed by a grant from the Medical Research Council.

MATERIAL.

Data of stature were furnished by seventeen mental hospitals included in the diet study and by another mental hospital not so included. Scrutiny of these data leads to the conclusion that they are reasonably accurate and provide a satisfactory basis of comparison. In the first mentioned seventeen Mental Hospitals the ages of the patients were also recorded.

Data of weights are more scanty and considerably less accurate. All the seventeen Mental Hospitals included in the diet study furnished statistics of (*a*) weight on admission, (*b*) weight at the date of making the record, but only six of these gave the weight without clothing and the estimates of the weight of clothing in the remainder were so conflicting that, save for one purpose, it was necessary to discard the statistics. The mental hospital not included in the diet study provided satisfactory records of weight, but not weight on admission.

In this paper we treat the subjects under the following headings:

1. Stature of the Mental Hospital Population Compared with Other Mental Hospital and Normal Populations.
2. Weights of Mental Hospital Patients Compared with the Weights of Normal Persons, taking account of Height and Age.
3. Changes of Weight in Inmates of Mental Hospitals During Residence.

STATURE.

Excellent material for comparison has been fully analysed by biometricians, notably Dr Tocher and the late Dr Charles Goring, whose work on the anthropometry of criminals is classical[*].

In Table I*a* we set out our own results and in Table I*b* those of previous students. It will be seen on comparison of our results with those of Tocher that the agreement is good. Tocher distinguished a "General Insane" from an "Entire Insane" population, the former derived from the latter by exclusion of individuals "considered by their medical attendants to have characters affected by special causes not characteristic of insanity in general" (Tocher, p. 301)[†]. We have excluded no cases, so that our sample approximates more closely to the "Entire Insane" than to the "General Insane" of Tocher's grouping.

TABLE I*a*.

Mean Height with Probable Errors and Standard Deviations.

Asylum		Males			Females		
		No. of observations	Mean Height	Standard Deviation	No. of observations	Mean Height	Standard Deviation
Northern	T	182	66·27 ± ·148	2·96	183	61·79 ± ·143	2·87
	S	133	65·97 ± ·137	2·34	—	—	—
	P	249	65·19 ± ·107	2·50	231	60·64 ± ·136	3·06
	M	175	65·91 ± ·144	2·82	189	60·04 ± ·157	3·20
	F	174	65·74 ± ·200	3·90	215	61·46 ± ·148	3·22
	G	146	65·70 ± ·123	2·21	183	62·27 ± ·143	2·87
Eastern	A	217	65·52 ± ·133	2·91	230	61·06 ± ·133	3·00
	B	149	65·35 ± ·183	3·30	195	61·04 ± ·145	3·00
	N	126	65·37 ± ·179	2·98	152	61·02 ± ·130	2·38
	O	63	64·96 ± ·248	2·92	60	64·02 ± ·164	1·88
Midland	E	39	64·25 ± ·519	4·80	83	61·42 ± ·193	2·60
	D	95	64·53 ± ·222	3·20	135	59·89 ± ·158	2·73
	R	244	64·98 ± ·109	2·52	250	61·15 ± ·142	3·33
South Western	K	61	65·00 ± ·195	2·25	95	62·06 ± ·233	3·37
	H	110	65·49 ± ·182	2·83	151	60·80 ± ·185	3·37
South Eastern	J	156	65·62 ± ·181	3·35	248	62·10 ± ·126	2·95
	L	152	66·02 ± ·172	3·15	132	61·45 ± ·164	2·79
All Asylums		2471	65·55 ± ·041	3·05	2732	61·27 ± ·040	3·10

---

[*] Goring, *The English Convict*. London, 1913 (Stationery Office).

[†] Tocher, "Anthropometry of Scottish Insane," *Biometrika*, Vol. v. 1906, p. 298.

TABLE I*b*.

Males. Mean Stature (in inches) and its Probable Error for Professional Classes, Air Force Candidates, Convicts, Insane, and certain Occupational Classes.

	Various classes in order of mean Heights	Commercial classes, Clerks and Shopkeepers	Selected classes, Soldiers, Policemen, Messengers, Servants	Labourers: Agricultural, Roads, Quarries, Railways	Sailors, Fishermen	Miners, Coal and Mineral	Artisans	Factory Operatives
Schuster's Results (Oxford Undergraduates, *Biometrika*, Vol. VIII)	69·49 ± ·057	—	—	—	—	—	—	—
British Association Report, 1881. Population, Ages 25–30	69·14 ± ·15	67·95 ± ·12	(	67·51 ± ·05		)	66·61 ± ·09	65·92 ± ·11
Macdonell's Cambridge Students (*Biometrika*, Vol. v. p. 347)	68·86 ± ·05	—	—	—	—	—	—	—
Pearson's English Sons (*Biometrika*, Vol. v. p. 347)	68·86 ± ·05	—	—	—	—	—	—	—
Air Force Candidates (Cripps, Greenwood and Newbold, *Biometrika*, Vol. XIV. p. 316)	68·83 ± ·048	—	—	—	—	—	—	—
Air Force Candidates passed Fit (Cripps, Greenwood and Newbold, *Biometrika*, Vol. XIV. p. 316)	68·82 ± ·054	—	—	—	—	—	—	—
Roberts, General Population, Ages 23–30	68·66 ± ·12	—	67·93 ± ·03	—	—	—	66·43 ± ·12	—
" " " " 23–50	67·93 ± ·15	67·28 ± ·10	67·22 ± ·02	67·10 ± ·04	66·37 ± ·09	66·91 ± ·19	66·77 ± ·09	66·89 ± ·14
Galton's S. Kensington Lab. Data, Pearson's *Chances of Death*, Vol. I. p. 311	(67·90 ± ·06	)	—	—	—	—	—	—
Pearson's English Fathers (*Biometrika*, Vol. v. p. 347)	67·74 ± ·05	—	—	—	—	—	—	—
Tocher's General Insane Population (*Biometrika*, Vol. v. p. 303)	65·9 ± ·029	—	—	—	—	—	—	—
Tocher's Entire Insane Population (*Biometrika*, Vol. v. p. 303)	65·7 ± ·029	—	—	—	—	—	—	—
English Mental Hospitals' Data	65·55 ± ·041	—	—	—	—	—	—	—
Bexley Mental Hospital	65·54 ± ·062	—	—	—	—	—	—	—
Convicts (Goring)	65·46 ± ·036	—	—	—	—	—	—	—
Tocher's Roxburgh and Selkirk Volunteers (*Biometrika*, Vol. v. p. 347)	—	—	67·89	—	—	—	—	—
Tocher's Aberdeenshire Rural Population (*Biometrika*, Vol. v. p. 347)	—	—	—	67·72	—	—	—	—
British Recruits (*Report on Health of the Army*, 1909)	—	—	66·22 ± ·01	—	—	—	—	—
British Association Report, 1881. All males over 25	67·92	—	—	—	—	—	—	—

It appears that the males of the English Mental Hospitals are very slightly shorter on the average and the females slightly taller than those of the Scottish Mental Hospitals.

Tocher inquired whether the differences between the means of the twenty-two Mental Hospitals supplying his data were consistent with random sampling, and concluded that the inter-local differences were significant.

We have used a method of analysis similar to that of Tocher (see Tocher, *op. cit.*, p. 315 *et seq.*) and likewise conclude that there is inter-local differentiation (Table II). In the lower half of the table we set out the distribution of differences

TABLE II.

Differences of Heights from General Mean.

Asylum		Males			Females		
		Difference	Probable error of difference	Ratio of S.D. of sampling to difference	Difference	Probable error of difference	Ratio of S.D. of sampling to difference
Northern	T	+ ·72	·1525	+3·41	+ ·52	·1545	+ 2·52
	S	+ ·42	·1783	+2·09	—	—	—
	P	- ·36	·1303	- 2·37	- ·63	·1375	- 3·27
	M	+ ·36	·1555	+1·74	- 1·22	·1521	- 5·45
	F	+ ·19	·1559	+ ·68	+ ·20	·1426	+ ·95
	G	+ ·15	·1702	+ ·82	+1·00	·1545	+ 4·85
Eastern	A	- ·03	·1396	- ·16	- ·20	·1378	- 1·05
	B	- ·20	·1685	- ·77	- ·23	·1497	- 1·11
	N	- ·18	·1832	- ·70	- ·25	·1696	- 1·30
	O	- ·59	·2591	- 1·62	+2·75	·2699	+11·23
Midland	E	- 1·30	·3293	- 1·71	+ ·15	·2294	+ ·53
	D	- 1·02	·2110	- 3·17	- 1·38	·1799	- 5·89
	R	- ·57	·1317	- 3·63	- ·12	·1322	- ·60
South Western	K	- ·55	·2633	- 1·91	+ ·79	·2145	+ 2·33
	H	- ·06	·1961	- ·23	- ·47	·1701	- 1·77
South Eastern	J	+ ·07	·1646	+ ·27	+ ·83	·1327	+ 4·63
	L	+ ·47	·1688	+1·90	+ ·18	·1820	+ ·76

Males			Females		
P.E.	Observed	Theoretical	P.E.	Observed	Theoretical
±1	7	8·5	±1	4	8·0
±1—2	5	5·5	±1—2	4	5·2
±2	5	2·9	±2—3	2	2·1
			±3—	6	·7
Total	17 $P = ·408$		Total	16 $P = ·040$	

as multiples of the Probable Errors in comparison with the frequencies of a Normal Distribution and the deduced values of P. The number of observations is, however, very small and little importance attaches to the values of P computed from so short a series.

The range of mean statures is from just over 66·0 inches to 64·25 in males, and from 64·0 to 59·9 in females.

The mean stature of male convicts is nearly the same as that of our male patients, and their range (see Goring, p. 195) similar to that found in the sample of Mental Hospitals.

The explanation tendered by Goring to account for the inferiority of the convicted criminal population in stature, viz. that they represented an unfavourable selection of the general population, may, we think, be applied with even greater plausibility to the case of inmates of mental hospitals, a proportion of whom are known to have suffered from defects of development.

The concordance of our general results with those of Tocher is such that we have little doubt in inferring that the constants here recorded do typify the "Entire Insane" population of English Mental Hospitals, and even less doubt in concluding that this population is physically select, its males being, on the average, shorter than their social and racial class. Of course the absolute difference may be small. There is (see Table I*b*) a wide range of stature in normal populations, and it is probable—but not certain—that the population from which a majority of patients in public mental hospitals are drawn approximates in characters more closely to the Artisan Class of Roberts' survey than to any other group.

The occupational distribution of the patients forming the sample analysed was not given, but the Board of Control supplied us with the classification of the whole mental hospital population for 1908–12, which was compared with that of the General Population shown in the census of 1901. In his study of criminals, Goring employed the occupational classification of Roberts* and used the latter's standard population, viz. that of the census of 1871. We have not found in Roberts' book any definition of his categories which would enable us to group the census returns of 1901 by his method, and we think that in any case 1871 is too remote from the epoch of our observations to be a very good standard. We have therefore re-grouped Goring's data (obtained from the schedule of measurements which was separately published) and set out in Table III the three proportional distributions.

Obviously neither the Convicts nor the Asylum Patients are random samples of the General Population. It must of course be remembered that a large discrepancy under headings XXII and XXIII would be expected, having regard to the fact that neither the Convict Prisons nor the Mental Hospitals will contain a normal proportion of persons between the ages of 10 and 20.

With reference to other groups showing wide differences, both Convicts and Mental Hospital patients show larger proportions of persons engaged in Professional Occupations and in National Defence, and both show smaller proportions of miners and of textile workers. The Convicts show a defect of agricultural workers and of persons employed in the food, tobacco and drink industries; in each of these cases the Mental Hospital patients show percentages nearer to the standard. On the other hand, the proportion of Mental Hospital patients drawn from the Building Trades differs more from the standard (in defect) than does that observed among Convicts.

---

* Roberts, *Manual of Anthropometry*. Churchill, London, 1878.

TABLE III.

Occupational Group (Males)		England and Wales Population Census 1901	Asylum Population 1908–12	Criminals (Goring) Population 1902-8
	Total population less children under 10 years	12,134,259	10,304	3,000
		Rates per 1000	Rates per 1000	Rates per 1000
I	General or Local Government ...	14·1	11·4	3·3
II	Defence of Country	13·9	28·2	27·7
III	Professional	25·7	39·9	48·2
IV	Domestic	25·1	29·6	27·7
V	Commercial	43·7	57·8	71·7
VI	Conveyance of men, etc. ...	102·9	93·0	90·7
VII	Agriculture	88·3	75·1	29·7
VIII	Fishing	2·0	2·3	6·7
IX	Mines	65·9	46·5	48·7
X	Metals, Machines, etc.	96·8	63·0	73·0
XI	Precious metals, jewels, etc. ...	10·8	6·9	17·0
XII	Building	85·9	69·3	87·2
XIII	Wood, Furniture, etc.	19·2	17·1	18·3
XIV	Brick, Cement, Pottery, etc. ...	11·7	4·9	5·7
XV	Chemical, Oils	8·4	4·8	5·0
XVI	Skin, Leather, etc.	6·6	5·8	3·7
XVII	Paper, Stationery	15·5	14·2	9·0
XVIII	Textiles	40·6	24·9	12·7
XIX	Dress	34·2	39·9	53·0
XX	Food, Tobacco, Drink	63·8	53·5	37·0
XXI	Gas, Water, etc.	5·9	4·4	3·3
XXII	Other general and undefined ...	56·0	200·2	285·7
XXIII	Without specified occupation or unoccupied	163·0	107·3	35·0

A perusal of this table does not suggest that the occupational groups manned by the middle and upper classes are under-represented in the Mental Hospital Population, and therefore does not offer an obvious explanation of any inferiority of physique observed. Thus the same conclusion as drawn by Goring (*op. cit.*, p. 178) holds, but less securely than in his case because we have not been able, as he was, to compare mental hospital patients and normal persons from the same occupational class. This is a task for the future, one thoroughly worth accomplishment.

In Table IV we give the correlation coefficients between stature and age, and the correlation ratios of stature upon age for normal and mental hospital populations. The differences are not such as to merit comment.

TABLE IV.

Correlation Coefficients and Ratios of Age and Height for Air Force and Insane.

	No. of Observations	Mean Age (Yrs.)	Standard Deviations	Mean Height (inches)	Standard Deviations	Correlation Coefficients	Correlation Ratios
Air Force Candidates. Males.	1238	23·92 ± ·097	5·07	68·83 ± ·048	2·50	− ·0339 ± ·019	·1790 ± ·019
English Mental Hospital. Males.	2471	41·94 ± ·209	15·41	65·55 ± ·041	3·05	·0092 ± ·014	·2298 ± ·013
English Mental Hospital. Females.	2732	43·68 ± ·199	15·42	61·27 ± ·040	3·10	− ·0777 ± ·013	·2654 ± ·012

BODY WEIGHT.

The subject of change of body weight with age and stature is evidently one of considerable importance. Its importance in *individual* prognosis is, of course, a commonplace of medical experience, since a deviation from the individual's norm affords valuable and early indications of disease. The value of a comparison of the weight of an individual with the standard provided by statistical averages of persons of his age, stature, race and social class is undoubtedly smaller, but the very extensive investigation published by the Medico-Actuarial Society of America* shows that even here importance can be attached to the procedure. It therefore seemed that results of interest would flow from an analysis of experience in Mental Hospitals. Unfortunately we have been hampered by two serious defects. The first is that data, even for normal populations, are less adequate than might have been expected; the second, that the measurements taken in the Mental Hospitals have been, on the whole, unsatisfactory.

The literature of adult body weight is far more scanty and much less exact than that of stature; the weight of adult women has indeed hardly been discussed at all in England and Wales during recent years. Most of the statistics of body weights relate to *clothed* persons, and the weights of the clothing have often been estimated in an unsatisfactory way. In the case of the Mental Hospitals, the estimates of weight of clothing had a range from 7 to 14 lbs., and we were quite unable to accept these estimates as having any pretensions to exactitude.

The best data regarding body weight in healthy adults to which we had access were measurements of 1238 Air Force candidates (Cripps, Greenwood and Newbold)†; these men were weighed stripped, and had a mean weight of 139·86 lbs. Goring's weights of 2500 convicts, weighed in trousers and shirt, are probably equally accurate. Their mean weight was 142·56 lbs., but their mean age was

---

* Medico-Actuarial Society of America. *Mortality Investigation. Statistics of Height and Weight of insured Persons.* New York, 1912, Vol. I.

† Cripps, Greenwood and Newbold, "A Biometric Study of the inter-relations of 'Vital Capacity,' Stature, Stem length and weight in a sample of Healthy Male Adults," *Biometrika*, Vol. XIV. Parts 3 and 4, 1923, p. 316.

36·24 years and the mean age of the Air Force candidates was only 23·92 years, while the statures differed by more than 3 inches. In general there is a positive correlation between weight and age, and between weight and height. Since the regression of weight upon age is not linear, the customary linear equations of regression are not exact methods of graduation, but they may be used to make rough allowance. Using the regression equation computed on the Air Force data, it appears that an Air Force candidate of the mean age and stature of the convicts would weigh 137 lbs. Allowing 2 lbs. (a minimum estimate) for the weight of trousers and shirt, we reach as the figure to be compared with Goring's, 139 lbs.

All other data relate to clothed men. The 3000 measurements made by Hutchinson* eighty years ago had a mean of 147·86 lbs. The 5032 men, aged 25 years or more, reported on by the British Association† forty years ago, had an average of 159·2 lbs. Lastly, Schuster's‡ weighings of undergraduates had a mean of 151·94 lbs. (these men were weighed without boots or coat).

Hutchinson's subjects included persons between 15 and 40, all but fifty-nine of Schuster's subjects were between 18 and 23, so that both samples might be expected to give lower averages than that of the British Association, which covered a larger proportion of older persons.

The Medico-Actuarial Society of America data refer to 221,694 men having a mean age of 33 years and a mean stature of 68·5 inches. The average weight—that of the clothed man—was 156·2 lbs. Levine§ has recently communicated the results of 28,697 examinations for life assurance in Great Britain and Ireland during the years 1921–22 (these, of course, are also weights of clothed men). From the graduated results, it appears that for a man of 36 standing 65½ inches (the mean values for Goring's convicts), the mean weight would be about 142 lbs. This is very near Goring's figure, and perhaps, since the candidates for assurance probably had more than shirt and trousers on, is evidence of a slightly lower mean weight.

Owing to the difficulty about clothing, we are really not in a position to state precisely what the average weight of a normal adult is. But, making allowance for stature, we think that the Air Force candidates were somewhat below the general average or, alternatively, that predictions based upon the Air Force data—involving the assumption of linear regression—slightly under-state the position, if the British Association data are taken as a standard. Thus the average age of the males over 25 in the British Association's experience was 34·9 years and the average stature 67·92 inches. Using the equation (deduced from the Air Force records)

$$\text{Weight (in pounds)} = 0{\cdot}7326 \text{ Age (in years)} + 3{\cdot}6133 \text{ Height (in inches)} - 126{\cdot}37,$$

---

* Hutchinson, "On the capacity of the lungs and on the Respiratory Functions with a view to establishing a precise and easy method of detecting disease by the Spirometer," *Transactions of the Royal Medico-Chirurgical Society of London*, Vol. XXIX.

† *British Association*, *Report* 1881. Report of the Anthropometric Committee, pp. 242–8.

‡ Schuster, "First Results from the Oxford Anthropometric Laboratory," *Biometrika*, Vol. VIII. 1911, p. 40.

§ Levine, "Heights and Weights," *Journal of the Institute of Actuaries*, Vol. LIV. Part 2, July 1928, p. 218.

we predict for this age and height, 144·6 lbs., or 14·6 lbs. below the observed weight. 14·6 lbs. seems an improbable figure for the weight of clothing. But on the Assurance Standard, the agreement is close. Levine's graduation would put the weight for this age and height at about 151 lbs., which only requires somewhat more than 6 lbs. to be allowed for clothing to coincide with the Air Force prediction. We do not of course know whether the older British Association figures or the modern assurance figures more correctly typify the normal average. If we accept the latter, there is very good concordance between the Air Force data and the assurance data, and, making allowance for difference of age and stature, between both and the weighings of criminals.

Upon the whole, therefore, we seem quite justified in judging our mental hospital data for males with the help of the equation deduced from the Air Force records.

Data of weights of women are, for England and Wales, extremely scanty. In Table V we give the results of the American Medico-Actuarial Investigation and three short series of English data. The difficulty about the allowance to be made for clothing is so great that any comparison is hazardous, but the following considerations suggest that the average weight of American women, allowance being made for height and age, is greater than that of English women. From the American data we have calculated the regression equation connecting weight, height and age and have applied this to the prediction of weights for the given mean ages and heights in three English series. We find for the Polytechnic women 128·1 lbs. compared with 124·4 observed, for the Somerset House women 125·6 against 117·2 observed, and for the Bedford College students 127·3 as against 117·8 observed. The Polytechnic Physical Training School women were as a whole particularly vigorous subjects and, bearing in mind both the difficulty about clothing and the fact that the range of age covered was very small, we cannot

TABLE V.

Means and Standard Deviations of Age, Weight, and Height for various groups of Women.

		Age (in years)			Weight (in lbs.)			Height (in inches)		
		Observations	Means	Standard Deviations	Observations	Means	Standard Deviations	Observations	Means	Standard Deviations
American Medico-Actuarial Investigation		136375	32·571	9·149	136375	134·196	19·428	136375	64·282	2·300
Cripps	Polytechnic Physical Training School...	232	20·52	2·29	226	124·38	11·95	226	64·78	1·58
	Somerset House ...	124	26·34	7·69	125	117·24	16·91	125	62·96	2·32
	Bedford College ...	115	21·13	1·79	70	117·75	16·29	70	64·42	2·52

Equation for Polytechnic Scholars: Weight = 3·9072 height − 128·73.

Equation for American Women: Weight = 3·4978 height + ·6437 age − 111·6156.

TABLE VI.

Comparison of Weights on Admission and at Time of Inquiry in Six Asylums where weights were made without clothes.

Asylum			No. of Observations	1st Weight (lbs.)		2nd Weight (lbs.)		Difference 1st & 2nd weight	Percentage who			Length of residence (months)	Age (years)	Calories (from dietaries)	Correlations of Height and Weight
				Mean	Standard Deviation	Mean	Standard Deviation		gained	lost	no change				
Northern	G	M	146	133·20 ± 1·06	19·05	138·81 ± 1·23	22·11	5·62	70	26	4	12·21	41·73	2926	+ ·5741 ± ·037
		F	183	115·52 ± 1·05	21·09	119·62 ± 1·09	21·92	4·10	58	30	12	10·72	47·28	2582	+ ·3730 ± ·043
	P	M	249	132·00 ± ·72	16·89	140·67 ± ·82	19·15	8·68	74	24	2	18·58	40·70	2249	+ ·4839 ± ·033
		F	231	107·46 ± ·96	21·55	115·38 ± 1·00	20·47	7·92	72	26	2	16·22	39·59	2006	+ ·2544 ± ·042
	T	M	182	121·05 ± ·92	18·40	125·47 ± ·94	18·79	4·42	62	34	4	12·83	43·03	2553	+ ·6173 ± ·031
		F	183	104·92 ± ·87	17·41	108·04 ± ·97	19·38	3·12	67	29	4	13·90	41·67	2218	+ ·2994 ± ·045
Eastern	N	M	126	133·27 ± 1·33	22·07	136·12 ± 1·33	22·10	2·86	59	39	2	12·02	43·20	3144	+ ·5119 ± ·044
		F	152	112·29 ± 1·27	23·21	112·26 ± 1·23	22·50	− ·03	53	45	1	12·39	44·25	2796	+ ·1994 ± ·053
Midland	D	M	95	118·62 ± 1·66	24·05	130·41 ± 1·62	23·39	11·79	83	16	1	12·20	43·24	2239	+ ·6251 ± ·042
		F	135	107·32 ± 1·15	19·87	109·99 ± 1·14	19·70	2·67	60	36	4	11·13	43·98	1902	+ ·4152 ± ·048
	R	M	244	124·55 ± ·83	19·29	132·79 ± ·82	19·09	8·24	76	21	3	13·68	38·82	2844	+ ·4630 ± ·034
		F	250	109·77 ± ·92	21·53	118·37 ± ·94	22·06	8·60	70	27	3	17·28	41·50	2565	+ ·3206 ± ·038
Six Asylums (Weighted Averages)		M	1042	127·44 ± ·40	19·29	134·43 ± ·44	20·95	6·98	71	26	3	14·16	41·34	2644	+ ·4991 ± ·016
		F	1134	109·49 ± ·42	20·83	114·48 ± ·43	21·47	4·99	65	31	4	14·07	42·74	2350	+ ·3099 ± ·018

be confident that their mean weight is really much less than that of the Americans. But the difference in both the other groups is so large that it seems most improbable that the true mean of either reaches the American Standard.

WEIGHTS OF MENTAL HOSPITAL PATIENTS.

In Table VI are given the particulars of weights for both males and females in six hospitals which supplied unclothed weights.

In Table VII we compare the weights of the males with the predictions from the Air Force equation, allowing for age and height.

TABLE VII.

Males. Means of Final Weights in Table IV compared with estimates from Air Force Equation.

Asylum		Observed weight	Predicted weight	Difference
Northern	G	138·8 ± 1·23	141·6 ± ·33	+ 2·8
	P	140·7 ± ·82	139·0 ± ·26	− 1·7
	T	125·5 ± ·94	144·7 ± ·30	+ 19·2
Eastern	N	136·1 ± 1·33	141·6 ± ·36	+ 5·5
Midland	D	130·4 ± 1·62	138·3 ± ·88	+ 7·8
	R	132·8 ± ·82	136·9 ± ·26	+ 4·1
Whole Sample		134·4 ± ·42	140·7 ± ·13	+ 6·3

The equation: Weight (lbs.) = ·7326 age (years) + 3·6133 height (inches) − 126·37.

In one instance, the observed weight exceeds the predicted weight, in the other five it is in defect.

Confining ourselves to persons over 20 years, the six mental hospitals provided 969 males and 1084 females. The biometric constants were evaluated and it appeared that the correlation of age with height was negligible. In Table VIII we show the mean weights for arrays of age together with the values assigned by the regression of weight upon age.

When these are tested in the usual way for Goodness of Fit* it appears—as indeed is obvious from the table—that the fit (in males) is bad, but that this is due to great divergence in particular groups. Thus, in the case of females, 38·6 °/₀ of the value of χ^2 is contributed by the last group, and its omission would increase P from ·25 to ·65 ($n' = 14$). In males the last two groups contributed nearly 62 °/₀ of the value of χ^2 and their omission improves P to ·073 (for $n' = 13$), and even in males the general trend is fairly well represented by a straight line. In other words, although the regression is certainly not linear, we may say that a comparison of the regression coefficients of weight upon age between equations derived from normal populations and from mental hospital populations is a justifiable proceeding. Such comparison shows us that the increase of weight with age is much smaller amongst the insane than amongst normal persons. While in respect of stature the insane are differentiated from the normal to the same extent as convicted criminals, the latter are not, in respect of weight, subnormal, the former are. The most important cause of this is, undoubtedly, that the population of a mental hospital contains many more physically diseased persons than a convict prison. Goring's analysis of the vital statistics of criminals showed *inter alia*, that "the current allegation of criminality and tubercular disease being kindred manifestations of the same form of human decadence is entirely unsupported by the statistical facts" (Goring, *op. cit.*, p. 229).

---

* *Biometrika*, Vol. ix. 1918, p. 28.

TABLE VIII.

Weights at different Ages for 969 *Males and* 1084 *Females in Six Asylums.*

Years of age	Weights of Males (lbs.)			Weights of Females (lbs.)		
	No. of Observations	Observed	Calculated	No. of Observations	Observed	Calculated
23	82	131·23	131·62	99	113·13	113·37
28	129	131·21	132·58	116	112·02	113·65
33	138	137·35	133·54	135	112·06	113·94
38	125	130·93	134·50	133	114 87	114·23
43	106	138·55	135·46	150	114·32	114·51
48	112	138·29	136·42	128	117·05	114·80
53	76	134·33	137·38	87	118·58	115·09
58	70	141·54	138·33	77	117·59	115·37
63	61	138·20	139·29	71	113·42	115·66
68	39	136·28	140·25	39	112·69	115·95
73	23	137·86	141·21	26	116·40	116·23
78	5	147·25	142·17	20	110·25	116·52
83	2	105·25	143·13	2	110·25	116·8
88	1	95·25	144·09	1	85·25	117·09

Equations: Weight $= \cdot 1917$ age $+ 127 \cdot 2156$ Weight $= \cdot 0573$ age $+ 112 \cdot 0487$

$n' = 15$ $n' = 15$

$x^2 = 50 \cdot 3658$ $x^2 = 17 \cdot 0964$

$P = \cdot 00$ $P = \cdot 25$

In mental hospitals, on the other hand, the rate of mortality and the prevalence of tuberculosis are very high. We cannot indeed account for our low average weights by the presence in the samples of overtly tuberculous patients, the numbers so recorded were very few, but that many were in a pre-tuberculous state can be regarded as probable. This, of course, at once raises the extremely interesting question, which is cause and which effect, but the present data are neither numerous enough nor accurate enough to allow us to discuss that question; a further peculiarity of the material is dealt with below. Our demonstration of the low weight for age and stature of the population is, we submit, a cogent argument in favour of requiring that changes of body weight in mental hospitals should be made the object of much more exact inquiry.

We are not acquainted with any published investigations of changes of weight in the chief clinical types of insanity and, having regard to the enormous field for investigation offered by the mental hospitals of this country, lack of such information cannot be justified. It appears that contempt for very humble methods of research, mere weighings and measurings, has been carried too far. Perhaps careful anthropometry would suggest to the clinician something of as much practical importance as is to be gathered from much recent psychological speculation.

TABLE IX.

Asylum			No. of Observations	Mean Age (years)	Mean Period of observation (months)	Mean Gain in weight (lbs.)	Percentage of patients observed who lost weight while under observation
Northern	F	M	174	42·1	11·8	6·54	20·7
		F	215	46·3	11·2	2·30	38·6
	G	M	146	41·7	12·2	3·66	26·0
		F	183	47·3	10·7	4·10	30·1
	M	M	175	43·5	13·9	4·27	28·0
		F	189	41·7	16·3	6·50	26·5
	P	M	249	40·7	18·6	8·21	25·7
		F	231	39·6	16·2	7·40	25·5
	S	M	133	43·4	14·5	4·31	31·6
		F	—	—	—	—	—
	T	M	182	43·0	12·8	3·97	33·5
		F	183	41·7	13·9	3·60	29·0
Eastern	A	M	217	41·7	12·3	5·49	30·4
		F	230	42·9	15·1	9·10	22·6
	B	M	149	42·1	13·7	7·28	22·1
		F	195	43·2	13·9	8·90	24·1
	N	M	126	43·2	12·0	3·39	38·9
		F	152	44·3	12·4	– ·20	45·4
	O	M	63	47·0	19·3	2·48	33·3
		F	60	45·3	13·9	4·90	28·3
Midland	D	M	95	43·2	12·2	11·15	15·8
		F	135	44·0	11·1	2·70	36·3
	E	M	39	48·2	13·3	11·00	17·9
		F	83	48·5	12·3	6·40	26·5
	R	M	244	38·8	13·7	7·95	21·3
		F	250	41·5	17·3	5·40	27·2
South Western	H	M	110	42·3	12·2	4·23	15·9
		F	151	45·8	10·5	5·40	18·5
	K	M	61	42·5	10·2	4·52	27·9
		F	95	44·9	11·7	7·00	24·2
South Eastern	J	M	156	41·4	11·6	3·27	34·6
		F	248	44·8	13·9	– ·60	46·8
	L	M	152	40·4	11·7	3·59	36·2
		F	132	44·3	13·1	6·10	27·3
Weighted Averages		M	2471	42·0	13·4	5·60	27·2
		F	2732	43·7	13·7	4·90	30·3

THE CHANGE OF WEIGHT OF PATIENTS BETWEEN ADMISSION AND DATE OF OBSERVATION.

In the above sections we have used the weights at the date of completing the records, and these were at various times after admission.

The subject of the change in weight of patients during residence in an asylum is a matter of considerable interest. In Tables IX and X we show the difference

TABLE X.

The Changes in Weight of Patients between Date of Admission and Date of Inquiry.

Mental Hospital			Total no. of observations in each asylum	Gained Weight				Lost Weight			
				No. who gained	Mean first weight	Mean gain	Mean final weight	No. who lost	Mean first weight	Mean loss	Mean final weight
Northern	f	M	174	130	122·9	11·28	134·2	36	132·0	8·28	123·7
	F	F	215	120	100·3	12·39	112·7	82	120·1	10·88	109·2
	g	M	146	102	132·0	11·37	143·4	38	138·9	7·23	131·7
	G	F	183	107	111·7	11·12	122·8	55	124·8	7·27	117·5
	m	M	175	112	131·9	11·21	143·1	50	145·6	11·46	134·1
	M	F	189	134	107·3	13·19	120·5	50	118·3	9·50	108·8
	p	M	249	185	129·7	14·36	144·1	59	140·9	8·97	131·9
	P	F	231	166	103·5	15·09	118·6	59	121·6	11·95	109·7
	s	M	133	82	130·2	9·18	139·4	43	132·5	5·07	127·4
	S	F	—	—	—	—	—	—	—	—	—
	t	M	182	112	117·3	10·80	128·1	61	128·5	6·59	121·9
	T	F	183	122	104·0	8·37	112·4	53	107·2	7·21	100·0
Eastern	a	M	217	140	121·4	11·80	133·2	66	133·1	6·28	126·8
	A	F	230	173	105·3	15·29	120·6	52	120·9	8·43	112·5
	b	M	149	109	125·1	12·89	138·0	33	138·1	8·09	130·0
	B	F	195	146	104·8	15·47	120·3	47	118 6	11·59	107·0
	n	M	126	74	128·2	12·74	140·9	49	141·4	10·89	130·5
	N	F	152	81	110·4	11·03	121·4	69	115·6	13·71	101·9
	o	M	63	37	128·8	10·22	139·0	21	142·6	8·19	134·4
	O	F	60	40	113·3	10·65	124·0	18	120·2	7·39	112·8
Midland	d	M	95	79	116·5	15·38	131·9	15	134·3	8·87	125·4
	D	F	135	81	102·3	13·14	115·4	49	114·6	13·12	101·5
	e	M	39	32	118·1	16·73	134·8	7	150·7	9·00	141·7
	E	F	83	59	112·9	12.89	125·8	22	129·9	9·68	120·2
	r	M	244	185	121·5	13·00	134·5	52	136·1	7·87	128·2
	R	F	250	175	106·3	15·69	122·0	67	120·5	9·23	111·3
South Western	h	M	110	74	127·5	11·11	138·6	29	133·4	10·69	122·7
	H	F	151	122	109·4	9·87	119·3	28	124·0	16·63	107·4
	k	M	61	41	133·3	10·48	143·8	17	143·9	12·29	131·6
	K	F	95	66	111·5	14·61	126·1	23	120·4	12·25	108·2
South Eastern	j	M	156	95	131·3	11·16	142·5	54	142·7	10·08	132·6
	J	F	248	117	108·6	10·54	119·1	115	123·8	10·93	112·9
	l	M	152	90	134·5	10·99	145·5	56	145·4	7·64	137·8
	L	F	132	92	105·6	12·59	118·2	35	110·6	8·37	102·2

between the weight on admission and the final weight in all the institutions making returns, including of course the six providing what we think are accurate weights. The others were included here because although, owing to the difficulty about the weight of clothing, the absolute values are useless, it is probable that the method of weighing in any one institution is constant so that the differences between initial and final weights may have some value. Even this, however, is open to question, because some of the weighings were in the summer and some in the winter, so that differences between summer and winter clothing are possibly of importance. There is, however, no uniformity in the seasonal distribution of initial and final weighings, so that such an error is randomly distributed.

It is very important to notice that the patients weighed are a peculiar selection of the inmates. The superintendents were asked (on May 30th, 1922) to supply the information* "concerning the first 250 male and 250 female patients admitted since January 1st, 1920, in order of their admission, who are still resident in your Mental Hospital. Should the number of admissions since the date mentioned be less than 250, the actual admissions will meet requirements." Consequently—the average time of residence of the patients measured being 13 months—there are excluded from the sample (*a*) persons who died within the average period of months from admission, (*b*) persons who were discharged cured within that period. We are therefore wholly dealing with persons suffering from some mental disorder of a relatively chronic type not accompanied at admission by bodily disease of a rapidly fatal type.

For the reasons stated we can, so far as the majority of the institutions are concerned, merely express the following *prima facie* deductions from Table X.

(1) There is no relation between scheduled diet and body weight.

(2) About 30 per cent. of the sampled population lose weight between admission and the date of final weighing, which is from 12 to 18 months after admission. The mean increases and decreases shown in Table X are in several instances large, but, bearing in mind the very serious criticism to which the methods of weighing are subject, no stress is put upon the precise figures; they do but indicate the need of a wider *and more accurate* survey.

(3) This result cannot be explained by the existence in the mental hospital population, as sampled, of a supra-normal proportion of persons suffering from bodily disease—such as phthisis—notoriously associated with loss of weight. Giving a very liberal interpretation to the terms physical disease likely to lead to loss of weight, we find that the proportion of patients in our sample belonging to this class is less than 5 per cent., and simple calculation shows that dilution of normal material with 5 per cent. of persons even enormously underweight would not bring down the general average as low as it is found to be.

(4) A plausible explanation is that many victims of acute psychoses are known to lose weight rapidly and that this sample will include a relatively large propor-

---

* **Extract of official letter of Board of Control.**

tion of such cases. On the other hand, the circumstances of persons of unsound mind coming from the economically less prosperous classes are often very unfavourable, and such persons would tend to put on weight when cared for in a public institution. These two considerations tend in opposite directions, qualitatively; what their respective values are we cannot tell.

For males of the average height of our asylum patients, viz. 65½ inches, the increase of weight for each quinquennium of age ranged from 3 to 5 lbs. between 45 and 20 and was rather less than 2 lbs. at older ages. We have seen that in the Air Force data the approximate increase for each year of age—stature constant—was rather less than ¾ lb. or somewhat more than 3½ lbs. for a quinquennium. This is a finding in good agreement with the Insurance Offices data.

Since the correlation of age and stature (adults) is very slight, the position is sufficiently well brought out by the accompanying Diagram 1 which shows the means of arrays of weights for quinquennial ages.

Weights at different Ages for Males and Females in six good Asylums.

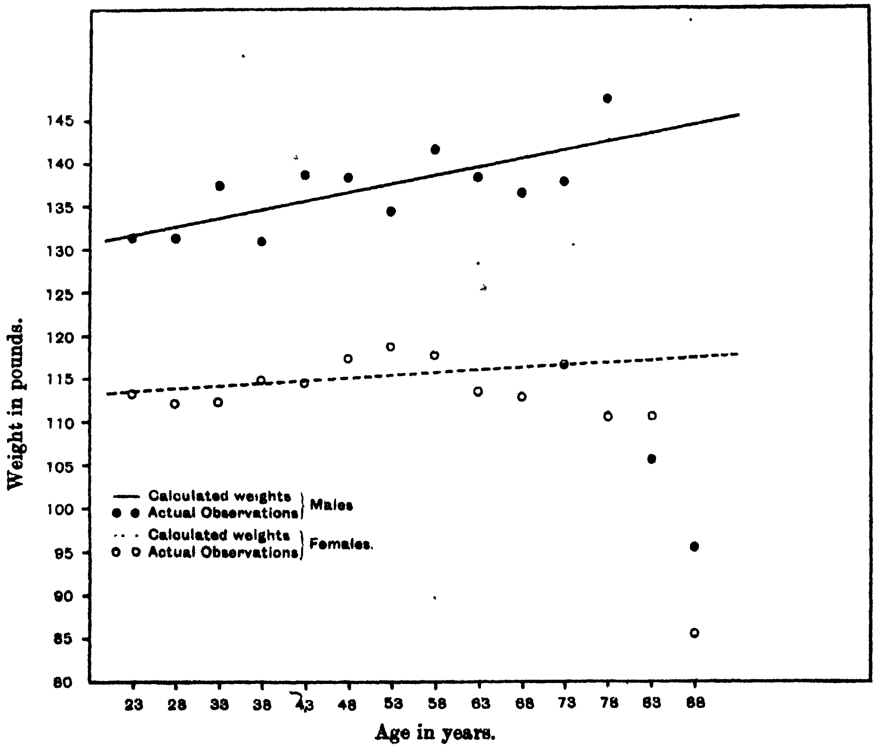


Diagram 1

Whether this divergence from the normal experience has a definite biological meaning or whether it is due to the unsatisfactory character of the material we are unable to say.

We think, however, that the results of this inquiry are sufficient to emphasize the need of a much more searching investigation of changes in weight of patients in mental hospitals than the present material has permitted us to make.

It is possible that the simplicity of the measurements has caused them to be despised as objects of scientific research, and certainly no one would claim for them a high rank in the appraisement of clinical conditions in individual prognosis or treatment. But it would, we think, be of some interest to ascertain whether the anthropometric characters of different clinical groups within the population of mental hospitals and the variations of such characters with age are significantly differentiated.

MISCELLANEA.

Some further Notes on Cancer and Goitre Distributions.

By PERCY STOCKS, M.D.

In a previous paper[*] evidence was collected which indicated a relation between the frequency of goitre and the incidence of cancer of the digestive organs, more especially the stomach. Since the publication of that paper some additional evidence of the same kind relating to the geographical distribution of these diseases in America and Sweden has been obtained, and is here set down.

I. *More Complete Analysis of American Statistics.*

During the Great War some $2\frac{1}{2}$ millions of recruits called up for service in the U.S. Army were subjected to a careful medical examination of which the records have been analysed in an official volume[†]. From these statistics the number of men per thousand examined from each State who exhibited various diseases can be ascertained. As regards anomalies of the thyroid gland the disorders included are "goitre (simple)," "exophthalmic goitre" and "cretinism." The rates for the last are too small to be reliable for the purpose in view.

The distribution of exophthalmic goitre is found to be highly correlated with that of simple goitre, the correlation coefficients between the frequencies in the various States being:

For the whole area ... ... $r = \cdot 674 \pm \cdot 054$
For 37 States east of Long. 104° W. ... $r = \cdot 830 \pm \cdot 031$
For 35 Registration States ... ... $r = \cdot 749 \pm \cdot 043$

This high positive correlation is difficult to reconcile with many accepted views as to the causation of Graves' disease, and is in contrast with the infrequency of this disease in goitrous areas of Switzerland. The question arises whether the dividing line between simple and exophthalmic goitres was the same in all the States, or whether in goitrous areas the Local Boards may not have looked more carefully for slight indications of hyperthyroidism and therefore classed a larger percentage of all goitres as exophthalmic than was the case elsewhere. To test this point we may separate the rejected men and those examined by Local Boards from the accepted men examined by Medical Boards in camps where the effect of any such local diagnostic peculiarity would be less: the correlations are, for rejected men $r = \cdot 611 \pm \cdot 062$, and for accepted men $r = \cdot 689 \pm \cdot 052$, from which it appears that the local factor as regards diagnosis is not of any importance. The correlation between Graves' disease and simple goitre distributions in the U.S.A. must therefore be accepted as a rather surprising fact, and it rather suggests that two distinct factors may be necessary to produce Graves' disease, viz. (i) a goitre-producing factor whose intensity is indicated by the prevalence of simple goitre, (ii) another factor, probably of nervous origin, which is only present to a slight extent in Switzerland, but is more pronounced in the U.S.A. and almost universally present in England and Wales. This, however, is mere speculation.

I have indicated in Table I and on Map No. 1 the total rate for simple goitre in each of the American States (i.e. number of men per 1000 examined by Medical Boards, whether rejected or accepted for service, in whom simple goitre was noted). In the Report of the 1900 Census the

---

[*] "Cancer and Goitre," *Biometrika*, Vol. xvi. Parts 3–4, Dec. 1924, pp. 364–401.

[†] "Defects found in Drafted Men." War Dept. Statistical Information, Washington; 1920.

TABLE I.

Goitre and Cancer Distribution in U.S.A.

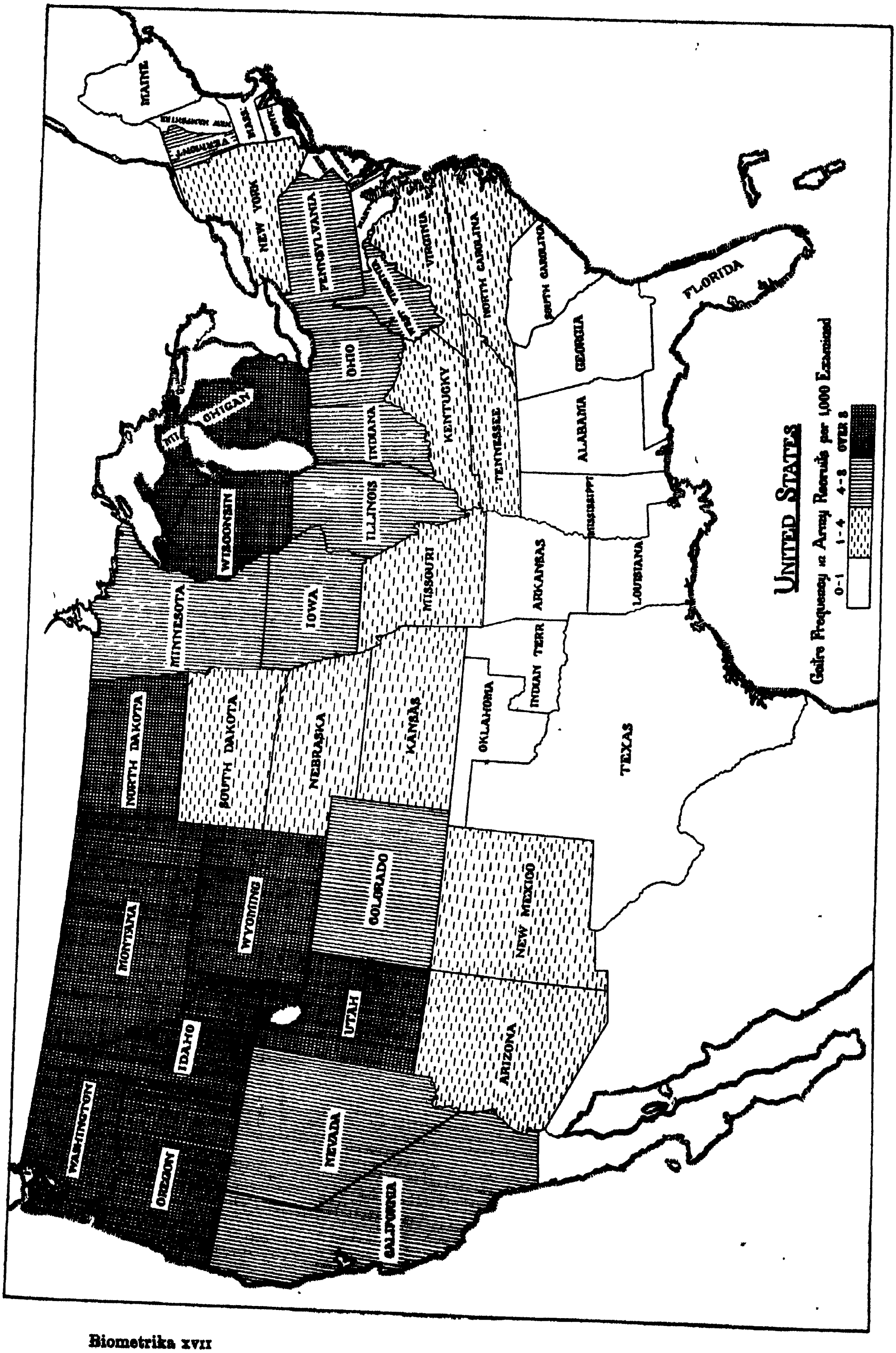
State	Simple Goitre per 1000 males (1)	Exophthalmic Goitre per 1000 males (1)	1900					1920				
			No. of deaths			Selection rate of Cancer for stomach (2)	Percent. males in population	Population	Percent. males in population	Per-cent. over 45	Corrected mortality rate	
			Cancer total	Cancer of stomach	Cancer, organ not stated						Cancer total (3)	Cancer stomach and liver (4)
Alabama	·56	·79	340	48	89	19·1	50·1	2,348,174	—		—	—
Arizona and New Mexico	1·07	·91	43	9	10	27·3	55·3	694,512	—	—	-	—
Arkansas	·40	·35	171	31	63	28·7	51·5	1,752,204	—	—	—	—
California	4·45	1·95	1026	295	271	37·7	55·2	3,426,861	52·9	26·1	92·2	42·7
Colorado	5·29	3·60	184	56	58	44·4	54·7	939,629	52·4	21·9	73·5	37·9
Connecticut ...	·89	1·53	555	95	158	23·9	50·0	1,380,631	50·4	22·2	96·1	44·6
Delaware	·59	·79	75	9	42	27·3	51·0	223,003	51·0	23·7	79·4	41·2
District of Columbia	1·39	3·47	—	—	—	—	—	437,571	46·5	22·8	85·9	40·5
Florida	·25	·66	83	19	38	42·2	52·1	968,470	51·1	19·4	55·8	23·9
Georgia	·52	1·61	373	30	141	12·9	49·8	2,895,832	—	—	—	—
Idaho	26·91	3·20	28	9	5	39·1	57·7	431,866	—	—	—	—
Illinois	7·79	5·57	2102	676	542	43·3	51·3	6,485,280	50·9	21·6	89·5	44·7
Indiana	6·49	3·12	1036	214	441	36·0	49·9	2,930,390	50·8	24·8	77·8	37·8
Iowa	6·68	2·86	801	273	253	49·8	51·8	2,404,021	—	—	—	—
Kansas	1·25	2·76	522	143	199	44·2	52·3	1,769,257	51·4	22·4	68·2	32·9
Kentucky	1·41	1·38	548	126	205	36·4	50·8	2,416,630	50·8	19·9	58·6	27·1
Louisiana	·62	·75	316	64	90	28·3	50·3	1,789,509	50·5	16·1	60·9	22·9
Maine	·66	·86	581	158	125	34·6	50·5	768,014	50·6	28·1	95·0	43·4
Maryland	·94	3·57	548	164	125	38·8	56·7	1,449,661	50·3	22·3	89·2	42·4
Massachusetts ...	·32	·63	1864	347	738	30·8	48·7	3,852,356	49·1	24·2	101·6	40·5
Michigan	11·43	6·89	1404	411	338	38·6	51·6	3,668,412	52·6	21·7	82·0	42·0
Minnesota	8·04	3·38	659	251	164	50·7	53·2	2,387,125	53·4	20·7	97·2	54·0
Mississippi ...	·64	·80	258	39	104	25·3	50·4	1,790,618	50·1	16·3	46·5	17·4
Missouri	3·99	4·80	908	229	251	34·9	51·4	3,404,055	50·6	23·1	71·1	35·2
Montana	21·00	3·39	55	22	7	45·8	61·6	548,889	54·6	18·4	57·5	27·2
Nebraska	2·14	1·70	260	82	144	45·8	52·9	1,296,372	51·9	20·5	78·2	43·1
Nevada	6·38	1·52	14	4	4	40·0	60·5	77,407	—	—	-	—
New Hampshire ...	·70	·94	275	46	156	38·7	49·9	443,083	50·1	28·5	90·0	39·1
New Jersey ...	·43	1·54	949	260	243	36·8	49·9	3,155,900	50·4	21·1	86·2	39·5
New York	1·19	2·94	4448	1097	1176	33·5	49·7	10,385,227	49·7	22·5	94·8	42·8
North Carolina ...	1·81	1·91	350	47	124	20·8	49·5	2,559,123	50·4	16·1	48·7	20·7
North Dakota ...	8·73	3·41	50	15	19	48·4	55·6	646,872	—	—	—	—
Ohio	5·59	5·08	1943	490	745	40·9	50·6	5,759,394	51·3	23·2	82·0	40·0
Oklahoma	·72	·87	49	14	12	37·8	53·8	2,028,283	—	—	—	—
Oregon	26·31	6·37	143	52	36	48·6	56·3	783,389	53·1	24·4	82·5	39·4
Pennsylvania ...	4·10	4·81	2607	688	746	36·9	50·9	8,720,017	50·8	20·9	83·5	41·9
Rhode Island ...	·55	1·24	267	57	44	25·5	49·1	604,397	49·2	23·3	90·7	37·1
South Carolina ...	·94	1·73	231	18	81	12·0	49·6	1,683,724	49·8	14·4	44·2	18·1
South Dakota ...	4·09	3·18	77	24	25	46·1	53·8	636,547	—	—	—	—
Tennessee	1·96	2·07	447	66	161	23·1	50·5	2,337,885	50·2	18·6	51·1	17·8
Texas	·30	·35	478	108	165	34·5	51·8	4,663,228	—	—	—	—
Utah	17·52	3·82	68	20	27	48·8	51·2	449,396	51·6	16·6	66·9	39·0
Vermont	2·14	4·28	281	60	115	36·1	50·9	352,428	50·8	28·5	96·4	43·9
Virginia	3·38	3·99	461	91	148	28·7	49·9	2,309,187	50·6	18·5	61·8	25·9
Washington ...	23·40	9·42	185	47	67	39·8	58·7	1,356,621	54·2	22·7	77·8	40·5
West Virginia ...	7·89	4·52	188	54	50	39·1	52·1	1,463,701	—	—	—	—
Wisconsin	14·02	7·94	980	313	390	53·1	51·6	2,632,067	51·5	22·1	85·0	50·1
Wyoming	15·38	2·71	16	6	5	54·5	62·9	194,402	—	—	—	—

(1) Total rate from all recruits examined, whether accepted or rejected.

(2) Corrected by omitting from total cancer deaths those where organ was unspecified.

(3) Corrected for age and sex distribution by approximate standardization from data in preceding columns.

(4) Corrected for cancers of unspecified site, and for age and sex by same methods as (3).


UNITED STATES
Goitre Frequency in Army Recruits per 1,000 Examined
0-1
1-4
4-8
OVER 8
MAINE
NEW HAMPSHIRE
VERMONT
MASS.
NEW YORK
PENNSYLVANIA
VIRGINIA
WEST VIRGINIA
NORTH CAROLINA
SOUTH CAROLINA
GEORGIA
FLORIDA
ALABAMA
MISSISSIPPI
LOUISIANA
TENNESSEE
KENTUCKY
OHIO
INDIANA
ILLINOIS
MICHIGAN
WISCONSIN
MINNESOTA
IOWA
MISSOURI
ARKANSAS
INDIAN TERR
OKLAHOMA
TEXAS
KANSAS
NEBRASKA
SOUTH DAKOTA
NORTH DAKOTA
MONTANA
WYOMING
COLORADO
NEW MEXICO
ARIZONA
UTAH
IDAHO
NEVADA
CALIFORNIA
OREGON
WASHINGTON

MAP No. 1.

number of deaths during that year from cancer of all organs and for cancer of the stomach were recorded for all the States, but in subsequent years mortality figures were confined to a limited number of "Registration States." In the 1920 report these number 35, and in these more recent records, cancer of the stomach and liver are grouped together. In 1900 the statistics for many of the sparsely populated Western States were of somewhat doubtful value and I have therefore correlated goitre frequency with the cancer factor in this year for (i) the whole area, and (ii) all States east of Long. 104° W., and for the year 1920 for (iii) the 35 Registration States.

The actual mortality rates from cancer depend so greatly on the proportions of negroes to whites in the populations of the various States that they do not afford a true index of the real incidence of the disease, but the selection rates of cancer for any one organ (i.e. ratio of cancer of that organ to cancer of all organs) are less liable to distortion from this cause. I have only attempted to use actual death-rates for correlation in the case of the 1920 mortality data.

A second source of error is the variation in sex-ratio in the populations of States, the north-western States being characterized by a preponderance of males. Since cancer of the stomach forms a larger percentage of male cancers than of female cancers, the selection rate of cancer for the stomach in the whole population is positively correlated with the proportion of males. The mortality statistics do not separate the sexes and I have therefore attempted to eliminate this source of error by partial correlation in the 1900 data, whilst the 1920 mortality rates for cancer generally and for cancer of stomach and liver have been independently corrected for differing sex-distribution in the usual way before correlating.

A third source of error, which may also affect the selection rate as well as the actual death-rates, is differing age distribution, and an approximate standardization of the deaths has been carried out for 1920, again treating cancer of all organs and cancer of stomach and liver separately. In this way a selection rate for each State was obtained for 1920 which had been corrected for unclassified cancer deaths and also in a rough manner for differing age and sex distribution. For this year it was also possible to apply Professor Pearson's method of correlating actual numbers of cases for constant population in order to eliminate possible spurious correlation.

The principal cancer statistics used for obtaining these correlations are given in Table I, and the results are given below. For the sake of comparison correlations have also been worked out with the total tuberculosis rate in recruits, the diabetes rate in recruits, the diabetes death-rate in 1920, and the exophthalmic goitre rate in recruits.

(α) For the whole area*, correlating goitre rate in men with selection rate of cancer for the stomach obtained from 1900 statistics (correcting for cancer deaths where the organ was unspecified)

$$r = \cdot 509 \pm \cdot 073, \qquad \eta_{c\,a} = \cdot 759.$$

This is clearly not a linear regression, but it becomes approximately so on omitting the 11 Western States with density of population less than 10 to the square mile, in which it is suspected that diagnosis was not so reliable (see β). The fact that the value of r then becomes larger suggests that there is a real relation which tends to be obscured by faulty statistics.

For comparison the following correlations over the same area may be noted.

Tuberculosis rate with goitre rate	$r = -\cdot 290 \pm \cdot 090$
" " with selection rate of cancer for stomach	$r = -\cdot 244 \pm \cdot 092$
Diabetes rate " " " " "	$r = \cdot 254 \pm \cdot 092$
" " " goitre rate	$r = \cdot 156 \pm \cdot 096$
" " " exophthalmic goitre rate	$r = \cdot 281 \pm \cdot 091$

(β) For the 37 States east of Long. 104° W., correlating goitre rate in men with selection rate of cancer for the stomach obtained from 1900 data as before

$$r = \cdot 644 \pm \cdot 065, \qquad \eta_{c.a} = \cdot 685.$$

---

* **47 areas, combining Arizona and New Mexico owing to their very small figures, and excluding Indian territory, Alaska and District of Columbia.**

Since the regression now approximates to linearity I have endeavoured to eliminate the effect of differing sex distribution by partial correlation; and also to test whether the exophthalmic goitre correlation with the cancer factor arises from its own correlation with simple goitre.

Selection rate of cancer for stomach with percentage of males in population	$r =$ ·534 ± ·079
Goitre rate with percentage of males … … … … … … … … … …	$r =$ ·219 ± ·105
Selection rate with goitre, percentage of males being constant … … …	$r =$ ·639 ± ·052
Selection rate with exophthalmic goitre rate … … … … … … … …	$r =$ ·568 ± ·075 $\eta_{C.E} =$ ·604
Goitre rate " " " " … … … … … … … …	$r =$ ·830 ± ·034 $\eta_{E.G} =$ ·896
Exophthalmic goitre rate with percentage of males … … … … …	$r =$ ·287 ± ·102
Selection rate with exophthalmic goitre rate, percentage of males and simple goitre rate being constant … … … … … … … … … … … …	$r = -$·031 ± ·111
Selection rate with goitre rate, percentage of males and exophthalmic goitre rate being constant … … … … … … … … … … … …	$r =$ ·447 ± ·089

The last two figures suggest that the correlation between exophthalmic goitre and the cancer selection rate arises entirely from the relation between exophthalmic and simple goitre distributions. For comparison the following have been computed, percentage of males being constant.

Tuberculosis rate with selection rate of cancer for stomach	$r = -$·268 ± ·103
Diabetes rate " " " " "	$r =$ ·215 ± ·106
" " " goitre rate … … … … … …	$r =$ ·168 ± ·108
" " " exophthalmic goitre rate … … …	$r =$ ·229 ± ·105

It may be noted that a positive relation between diabetes and total cancer death-rates for the American cities was demonstrated by Dr Maynard in this Journal (Vol. VII. p. 276)*. A significant relation between diabetes and exophthalmic goitre would be of interest in view of recent claims as to the beneficial effect of insulin treatment in Graves' Disease. If the diabetes death-rates for 1920 are correlated with those of exophthalmic goitre, however, $r =$ ·148 ± ·108, and with those of simple goitre $r =$ ·025 ± ·111, neither being significant.

(γ) For the 35 Registration States, correlating goitre rate in men with cancer mortality statistics for 1920, after correcting for age and sex distributions as well as for cancers of unspecified organ.

(i) Goitre cases, estimated, with deaths from cancer of stomach and liver, corrected … … … … … … … … … … … … … … … …	$r =$ ·578 ± ·076
Ditto for population constant … … … … … … … … … … … …	$r =$ ·438 ± ·092
(ii) Goitre rate with corrected mortality rate from cancer of stomach and liver for total cancer death-rate constant … … … … … … … … …	$r =$ ·516 ± ·084
(iii) Goitre rate with selection rate of cancer for stomach and liver … …	$r =$ ·461 ± ·090

It will be noticed that the three methods of correlation give results of the same order.

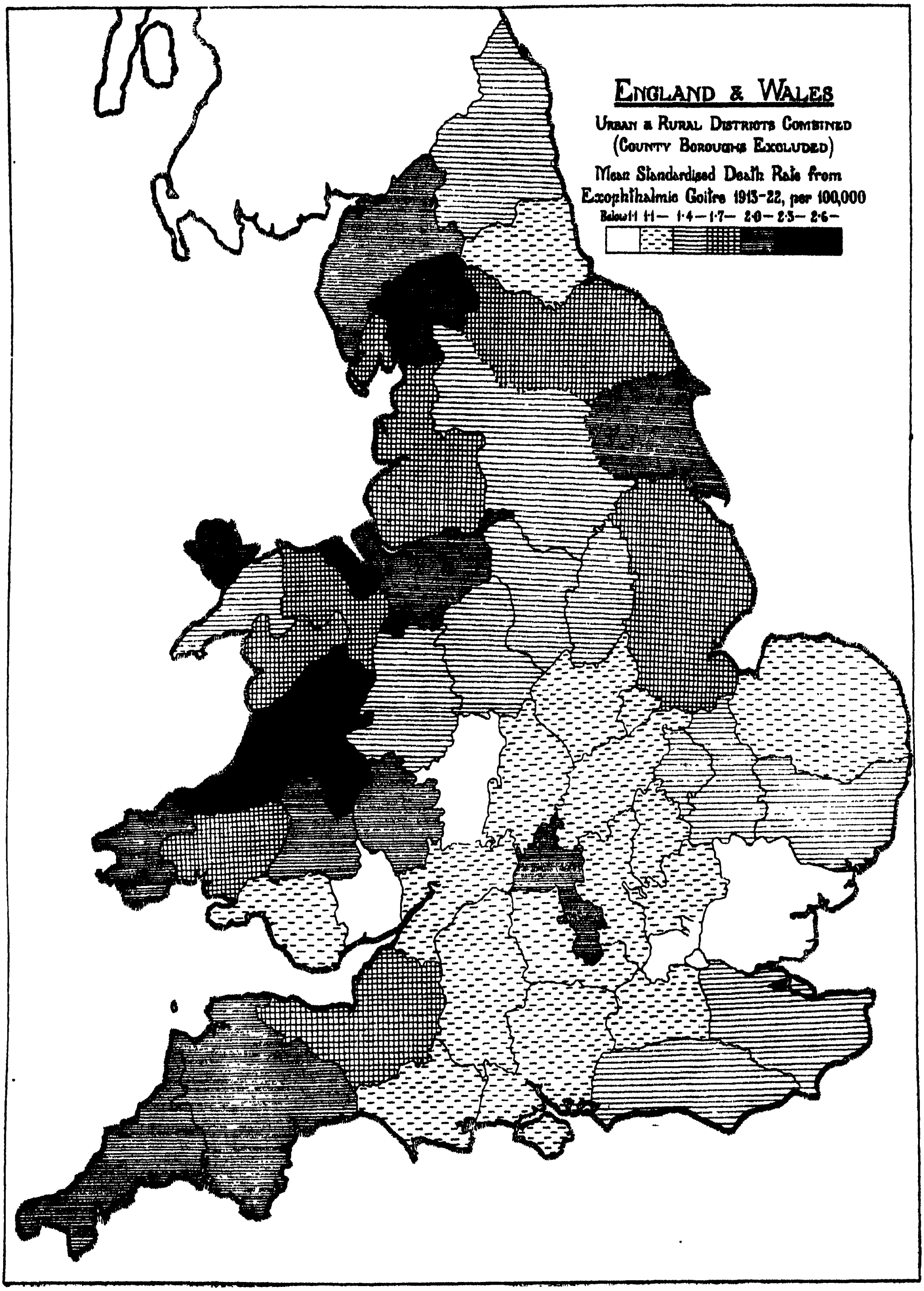
These figures seem to indicate that there is evident in the American States as in Switzerland a significant relation between goitre frequency and the tendency of cancer to attack the stomach. In view of the more complex factors involved and the dubious value of the mortality statistics for some of the States this conclusion is put forward with due reserve, but the correspondence between the results obtained in such widely differing countries is remarkable.

II. *Goitre and Cancer in Sweden.*

Through the kind cooperation of Mr R. Moosberg of Stockholm I have been able to obtain statistics of investigations on goitre frequency in the two counties where goitre is most prevalent. In 1892 a survey of school-children aged 8–16, of confirmees aged 16–18 and of recruits for military service aged 21 and over was carried out in the Departments of Kopparberg and Gefleborg. In the former county as a whole 346 children out of 2221 examined were found to be

---

* For a criticism of Maynard's result: see *Biometrika*, Vol. XI. p. 191 *et seq.*


ENGLAND & WALES
URBAN & RURAL DISTRICTS COMBINED
(COUNTY BOROUGHS EXCLUDED)
Mean Standardised Death Rate from
Exophthalmic Goitre 1913-22, per 100,000
Below 1·1
1·1—
1·4—
1·7—
2·0—
2·3—
2·6—

MAP No. 2,

goitrous (15·6 %), 309 confirmees out of 1624 (19 %) and 109 soldiers out of 677 (16 %). A more recent survey in 1910 of the district of Husby in the same county showed 15·9 % of males and 33·6 % of females to have goitre. Other important endemic areas in the county were the districts of Bjursås, Sundborn, Stora Tuna and Svärisjo. In the Department of Gefleborg goitre was found to be endemic in a few areas, but for the county as a whole the percentage of recruits affected was only 1·3 % and of children 12·2 %. The most goitrous districts were Ockelbo with 44 % of children and 30 % of adolescents goitrous, and Bjenåker and Alfta with lower rates. A later investigation in the Högbo district showed 5 % of men and 30 % of women to be goitrous.

I have shown in Table II the total cancer deaths, deaths from all causes, deaths from unknown causes and corrected cancer mortality rates for the year 1911 for (i) the whole county of Kopparberg, (ii) goitrous areas of Ockelbo, Bjenåker and Alfta of the county of Gefleborg, and (iii) rural areas of Sweden as a whole.

TABLE II.

Area	Population 1911	Cancer deaths 1911	All deaths 1911	Deaths from unknown cause 1911	Corrected cancer death-rate
Department of Kopparberg ...	235,427	240	3067	240	—
Goitrous districts of Gefleborg	19,159	20	237	22	—
Total goitre areas	254,586	260	3304	262	110·9
Sweden (rural)	4,159,216	—	—	—	96·7

It will be seen that the goitrous areas had apparently a somewhat higher cancer mortality rate than rural Sweden as a whole. It is unfortunately not possible to obtain more extended data.

III. *Exophthalmic Goitre in England and Wales.*

In view of the interest which is at the present time being aroused on the question of goitre in England and Wales, I have thought it advisable to include with this paper a map (No. 2) showing the standardized mean mortality rates from exophthalmic goitre over the 10 years 1913–1922 in the various counties (excluding county boroughs) which has been compiled from the statistics of the Registrar-General. For explanation of the method used in standardizing reference should be made to Section VIII of the original paper in *Biometrika*, Vol. XVI. pp. 392—398, Dec. 1924.

Recurrence Formulae for the Moments of the Point Binomial.

By RAGNAR FRISCH, Kristiania.

Introduction. In *Biometrika*, May 1924, Professor Pearson has given a very important recurrence formula for the moments of the hypergeometric series. In the special case of the point binomial $(p+q)^r$ Pearson's formula may be written

$$\mu_s = rqp \sum_{i=0}^{s-2} \binom{s-1}{i} \mu_i - q \sum_{i=0}^{s-2} \binom{s-1}{i} \mu_{i+1} \quad \text{.....(1)}$$

or

$$\mu_s + q\binom{s-1}{1}\mu_{s-1} + \sum_{i=2}^{s} \left[q\binom{s-1}{i} - rqp\binom{s-1}{i-1} \right] \mu_{s-i} = 0 \quad \text{.....(2),}$$

where the moment of order s is defined to be

$$\mu_s = \mu_s(q) = \sum^{r} (\nu - rq)^s \, T_\nu,$$

where

$$T_\nu = \binom{r}{\nu} q^\nu p^{r-\nu}, \quad p+q=1 \qquad \text{......(3)},$$

$$\mu_0 = 1,$$

$$\mu_1 = 0.$$

In the following lines I first deduce some general properties of linear equations between the moments of the point binomial. I then proceed to prove a system of recurrence formulae, a special case of which is the formula (1). Lastly I generalize the recurrence formulae of Pearson and Romanovsky[*] for the case of incomplete moments. The proofs of the last-named generalizations are based on principles entirely different from those originally used by Pearson and Romanovsky. In fact only elementary summation operations are involved. By the help of these operations the proofs may be given an extremely simple form. The formula of Romanovsky in fact arises with hardly any calculation at all.

§ 1. *Some General Properties of Linear Equations between the Moments of the Point Binomial.* A function $F(x)$ which identically in x satisfies one of the equations

$$F(1-x) = +F(x) \qquad \text{......(4a)},$$

$$F(1-x) = -F(x) \qquad \text{......(4b)},$$

may be called a *reduced* function of x, *positively* reduced if it satisfies (4a), *negatively* reduced if it satisfies (4b). A positively reduced function obviously is an even function in the variable $t=1-2x$, and a negatively reduced function is an odd function in the same variable.

Putting $p=1-q$ for q in (3) we obtain

$$\mu_s(1-q) = \sum_{\nu=0}^{r} (\nu - r + rq)^s \binom{r}{\nu} (1-q)^\nu q^{r-\nu} = (-1)^s \sum_{\nu=0}^{r} ((r-\nu) - rq)^s \binom{r}{r-\nu} q^{r-\nu} (1-q)^{r-(r-\nu)},$$

hence

$$\mu_s(1-q) = (-1)^s \mu_s(q) \qquad \text{......(5)}.$$

The even moments are positively reduced functions of q and the odd moments negatively reduced functions of q. That is: the even moments are even functions and the odd moments odd functions in the variable $\delta = p - q = 1 - 2q$.

Now let us consider a linear relation between the $(a+1)$ moments

$$\mu_s, \ \mu_{s-1}, \ \dots \ \mu_{s-a},$$

$$\sum_{i=0}^{a} A_i(q) \, \mu_{s-i} = 0 \qquad \text{......(6)},$$

where the coefficients $A_i(q)$ are functions of q.

The function $\phi_i(q) = (-1)^i \cdot \dfrac{A_i(1-q)}{A_i(q)}$ may be called the *form-function* of the equation. If the form-function is independent of the subscript i, that is to say if a function $\phi(q)$ exists so that identically in q we have: $(-1)^i \cdot \dfrac{A_i(1-q)}{A_i(q)} = \phi(q)$ for all values $i=0, 1, \dots a$, then I call (6) a *reduced* equation If no such function exists, I call (6) a *not-reduced* equation.

The character of an equation as reduced or not-reduced evidently remains unaltered if the equation is multiplied by an arbitrary function of q.

---

[*] *Biometrika*, Vol. xv. p. 410, 1923. After finishing the present paper I received a letter from Professor Tchouproff who drew my attention to the fact that Romanovsky's formula was first found by Bohlmann. Cf. Bortkiewicz, *Jahresbericht der deutschen Mathematikervereinigung*, Bd. xxvii. (1918), S. 73. [Romanovsky's formula was communicated by him in 1915 to the Society of Naturalists of the University of Warsaw. Ed.]

In the special case $\phi(q) = \pm 1$ I call (6) a *completely reduced* equation, *positively* reduced if $\phi(q) = +1$, *negatively* reduced if $\phi(q) = -1$. A reduced equation which is not completely reduced is obviously transformed into a completely reduced equation when it is divided by any of its coefficients or any other function $F(q)$ which satisfies one of the equations

$$\phi(q) \,.\, F(q) = \pm F(1-q),$$

where $\phi(q)$ is the form-function of the given equation.

Now suppose that we have given a not-reduced equation of the form (6). Since this equation holds good for every value of the variable between 0 and 1, it holds good for the value $(1-q)$. So we have

$$\sum_{i=0}^{a} A_i(1-q)\,\mu_{s-i}(1-q) = 0.$$

Substituting the expression for $\mu_{s-i}(1-q)$ from (5) we get

$$\sum_{i=0}^{a} (-1)^i A_i(1-q)\,\mu_{s-i} = 0 \quad\text{..(7).}$$

The last equation evidently is a not-reduced equation. Furthermore the two equations (6) and (7) are linearly independent since no function $\phi(q)$ exists, so that for all the values $i = 0, 1, \dots a$ the coefficients $(-1)^i A_i(1-q)$ of the new equation (7) are $\phi(q)$ multiplied by the coefficients of the given equation (6).

Thus we have the following fundamental property of the moment equations: *To every not-reduced linear equation between the moments corresponds another not-reduced equation which is linearly independent of the first, and which may be written down immediately when the coefficients of the first equation are known.*

It is easily seen that this property does not belong to reduced equations. From the two not-reduced equations (6) and (7) we may by linear combination deduce an infinity of new equations. Of these however not more than two can be linearly independent. May it in this way be possible to obtain two linearly independent *reduced* equations?

Evidently it does not restrain generality if we suppose the equations to be obtained by adding to the equation (7) as it stands the equation (6) multiplied by some function $\theta(q)$ of q. This gives

$$\sum_{i=0}^{a} \left[\theta(q) A_i(q) + (-1)^i A_i(1-q)\right] \mu_{s-i} = 0 \quad\text{..............................(8).}$$

The form-function of this equation is

$$\Phi_i(q) = \frac{\theta(1-q)\,\phi_i(q) + 1}{\phi_i(q) + \theta(q)} = \theta(1-q)\,\frac{\phi_i(q) + \dfrac{1}{\theta(1-q)}}{\phi_i(q) + \theta(q)},$$

where $\phi_i(q)$ is the form-function of the given equation (6).

The necessary and sufficient condition that there exists a function $\Phi(q)$ independent of the subscript i, so that identically in q

$$\Phi_i(q) = \Phi(q)$$

for all the values $i = 0, 1, \dots a$, is that

$$\theta(q)\,\theta(1-q) = 1 \quad\text{..(8 b),}$$

for the values of q in the interval $0 < q < 1$ for which the form-function of (6), $\phi_i(q)$, is not independent of the subscript i. For the values of q for which (8 b) is satisfied, we have

$$\Phi(q) = \theta(1-q) = \frac{1}{\theta(q)}.$$

Hence the necessary and sufficient condition that (8) is a reduced equation is that $\theta(q)$ is a function with the specified properties. Furthermore it is evident that if we choose two such functions $\theta_1(q)$ and $\theta_2(q)$ so that $\theta_1(q) \neq \theta_2(q)$, the corresponding two equations of the form (8) are linearly independent.

So we have the second property of the linear equations between the moments: *Any not-reduced equation of the form* (6) *may be replaced by two reduced linearly independent equations of the form*

$$\sum_{i=0}^{a} [\theta_1(q) A_i(q) + (-1)^i A_i(1-q)] \mu_{s-i} = 0 \quad \text{..........................(9}a\text{)},$$

$$\sum_{i=0}^{a} [\theta_2(q) A_i(q) + (-1)^i A_i(1-q)] \mu_{s-i} = 0 \quad \text{..........................(9}b\text{)},$$

where $\theta_1(q)$ *and* $\theta_2(q)$ *are two arbitrary not identical functions of* q *satisfying the conditions specified under* (8 *b*).

In treating linear equations between the moments one will nearly always find it easier to handle the reduced than the not-reduced forms.

Application. The equation (2) is a not-reduced equation with coefficients

$$\begin{aligned} A_0 &= 1, \\ A_1(q) &= q\binom{s-1}{1}, \\ A_i(q) &= q\binom{s-1}{i} - rqp\binom{s-1}{i-1}, \quad i=2,3,\ldots s. \end{aligned}$$

Choosing $\theta_1(q) = +1$, $\theta_2(q) = -1$ we may replace the given equation by the following two reduced equations which are written down immediately by the help of (9*a*) and (9*b*),

$$2\mu_s = \sum_{i=1}^{\text{not v}} \left[-(q+(-1)^i p)\binom{s-1}{i} + (1+(-1)^i)\binom{s-1}{i-1} rqp \right] \mu_{s-i} \quad \text{...........(10}a\text{)},$$

$$\binom{s}{1}\mu_s = \sum_{i=1}^{\text{not v.}} \left[-(q+(-1)^i p)\binom{s}{i+1} + (1+(-1)^i)\binom{s}{i} rqp \right] \mu_{s-i} \quad \text{..............(10}b\text{)},$$

where the summation is to be continued to the last not vanishing ("not v.") term. The equation (10*b*) is the equation obtained from (9*b*) when $(s-1)$ is replaced by s.

Writing out the formula (10*a*) we get

$$\begin{aligned} \mu_s = {} & \left(\frac{p-q}{2}\right)\left[\binom{s-1}{1}\mu_{s-1} + \binom{s-1}{3}\mu_{s-3} + \ldots\right] \\ & + rqp\left[\binom{s-1}{1}\mu_{s-2} + \binom{s-1}{3}\mu_{s-4} + \ldots\right] - \frac{1}{2}\left[\binom{s-1}{2}\mu_{s-2} + \binom{s-1}{4}\mu_{s-4} + \ldots\right] \quad \text{......(11)}, \end{aligned}$$

where each of the sequences is to be continued to the last not vanishing term.

The formula (11) may be slightly more convenient for numerical calculations than the formula (1). Other formulae involving still less calculation may however be obtained as I shall show in the next paragraph.

§ 2. *A general System of Recurrence Formulae.* Let us denote by H_s the left-hand side of Pearson's equation (2),

$$H_s = \mu_s + q\binom{s-1}{1}\mu_{s-1} + \sum_{i=2}^{s}\left[q\binom{s-1}{i} - rqp\binom{s-1}{i-1}\right]\mu_{s-i}.$$

Calculating the first few differences of H_s,

$$\begin{aligned} \Delta H_s &= H_{s+1} - H_s, \\ \Delta^2 H_s &= \Delta H_{s+1} - \Delta H_s, \end{aligned}$$

the following formula for $\Delta^k H_s$ is suggested,

$$\Delta^k H_s = -\sum_{i=0}^{\text{not v.}} \left[(-1)^{i-1}\binom{k}{i} p - \binom{s-1}{i} q + \left(\binom{s-1}{i-1} + (-1)^i\binom{k}{i-1}\right) rqp \right] \mu_{s+k-i} \ldots\text{(12)}.$$

The formula may easily be proved by complete induction. It obviously holds good for $k=0$. Supposing it to hold good for k we have

$$\begin{aligned}
\Delta^{k+1} H_s &= \Delta^k H_{s+1} - \Delta^k H_s \\
&= -\sum_{i=0}^{\text{not v.}} \left[(-1)^i \binom{k}{i} p - \binom{s}{i} q + \left(\binom{s}{i-1} + (-1)^i \binom{k}{i-1}\right) rqp\right] \mu_{s+1+k-i} \\
&\quad + \sum_{i=0}^{\text{not v.}} \left[(-1)^{i-2} \binom{k}{i-1} p - \binom{s-1}{i-1} q + \left(\binom{s-1}{i-2} + (-1)^{i-1} \binom{k}{i-2}\right) rqp\right] \mu_{s+k-(i-1)} \\
&= -\sum_{i=0}^{\text{not v.}} \left\{\left[(-1)^{i-1} \left(\binom{k}{i} + \binom{k}{i-1}\right) p - \left(\binom{s}{i} - \binom{s-1}{i-1}\right) q\right]\right. \\
&\qquad \left. + \left[\binom{s}{i-1} - \binom{s-1}{i-2} + (-1)^i \left(\binom{k}{i-1} + \binom{k}{i-2}\right)\right] rqp\right\} \mu_{s+(k+1)-i} \\
&= -\sum_{i=0}^{\text{not v.}} \left[(-1)^{i-1} \binom{k+1}{i} p - \binom{s-1}{i} q + \left(\binom{s-1}{i-1} + (-1)^i \binom{k+1}{i-1}\right) rqp\right] \mu_{s+(k+1)-i},
\end{aligned}$$

which is the expression for $\Delta^{k+1} H_s$ obtained from (12).

From Pearson's equation we know that $H_s=0$ identically in s, consequently $\Delta^k H_s=0$.

Hence replacing $(s+k)$ by s in (12) we have the following equation,

$$\sum_{i=0}^{\text{not v.}} \left[(-1)^{i-1} \binom{k}{i} p - \binom{s-k-1}{i} q + \left(\binom{s-k-1}{i-1} + (-1)^i \binom{k}{i-1}\right) rqp\right] \mu_{s-i} = 0 \quad \ldots\ldots(13a),$$

where k denotes an arbitrary not negative integer. This is a not-reduced equation, which is easily seen by an inspection of the form-function of the equation*. Using the result of the preceding paragraph we may immediately write down the following not-reduced equation,

$$\sum_{i=0}^{\text{not v.}} \left[(-1)^{i-1} \binom{s-k-1}{i} p - \binom{k}{i} q + \left((-1)^i \binom{s-k-1}{i-1} + \binom{k}{i-1}\right) rqp\right] \mu_{s-i} = 0 \ldots(13b).$$

Taking first the sum and then the difference of the last two equations, which corresponds to choosing the functions $\theta_1(q)$ and $\theta_2(q)$ from the preceding paragraph equal to $+1$ and -1 respectively, we obtain the following two *reduced* equations,

$$\sum_{i=0}^{\text{not v.}} \left[\left(\binom{s-k-1}{i} + \binom{k}{i}\right) ((-1)^{i-1} p - q) + (1+(-1)^i) \left(\binom{s-k-1}{i-1} + \binom{k}{i-1}\right) rqp\right] \mu_{s-i} = 0 \quad \ldots\ldots\ldots\ldots\ldots(14a),$$

$$\sum_{i=0}^{\text{not v}} \left[\left(\binom{s-k}{i+1} - \binom{k}{i+1}\right) ((-1)^{i-1} p - q) + (1+(-1)^i) \left(\binom{s-k}{i} - \binom{k}{i}\right) rqp\right] \mu_{s-i} = 0 \quad \ldots(14b),$$

where k again denotes an arbitrary not negative integer, not necessarily the same in (14a) as in (14b). The equation (14b) is the equation obtained by subtracting (13a) from (13b) and replacing $(s-1)$ by s.

To every not negative value of k correspond two recurrence formulae (14a) and (14b). Putting $k=0$ we obtain the formulae (10a) and (10b). In this special case the formulae contain all the moments from μ_0 to μ_s. By choosing k conveniently we may reduce the number of moments occurring in the formulae. In fact if we choose any particular value $k_0 \left(<\frac{s}{2}\right)$ of k, the coefficients of all the moments from order zero to order $\begin{Bmatrix} k_0 - 1 \text{ (if } (s-k_0) \text{ be even)} \\ k_0 \text{ (if } (s-k_0) \text{ be odd)} \end{Bmatrix}$ vanish in the formula. If s be even and $k=\frac{s}{2}$, the coefficients in (14a) vanish for the moments from order zero to order $\begin{Bmatrix} k-1 \text{ (if } \tfrac{1}{2}s \text{ be even)} \\ k-2 \text{ (if } \tfrac{1}{2}s \text{ be odd)} \end{Bmatrix}$. If s be odd we may for instance choose $k=\frac{1}{2}(s-1)$†. Formula

---

* The only case in which (13 a) might be a reduced equation is the case $k=\frac{1}{2}(s-1)$.

† $k=\frac{1}{2}(s-1)$ being the case in which (13 a) is a reduced equation, the special formula (15), p. 170, may of course also be derived directly from (13 a) or from (13 b).

(14*a*) then gives

$$\mu_s=(p-q)\left[\binom{k}{1}\mu_{s-1}+\binom{k}{3}\mu_{s-3}+\dots\right]+2rqp\left[\binom{k}{1}\mu_{s-2}+\binom{k}{3}\mu_{s-4}+\dots\right]$$
$$-\left[\binom{k}{2}\mu_{s-2}+\binom{k}{4}\mu_{s-4}+\dots\right] \quad \dots\dots\dots\dots(15).$$

This formula involves considerably less calculation than the analogous formula (11). For instance to calculate μ_{13} by (11) we must use the values of all the moments from μ_0 to μ_{12} with numerical coefficients such as 495, 792, 924, etc. To calculate μ_{13} by (15) we only use the moments from μ_7 to μ_{12} with numerical coefficients 6, 15, and 20.

I think it should be possible by further transformation of Pearson's equation, using the results from § 1, to obtain an equation only containing four consecutive moments, but I have not yet been able to carry out this transformation.

§ 3. *Generalization of Pearson's and Romanovsky's Formulae to incomplete Moments.* The incomplete moment of order s may be defined as

$$\mu_s=\sum_{\nu=\rho}^{r}(\nu-rq)^s T_\nu,$$

where

$$T_\nu=\binom{r}{\nu}q^\nu p^{r-\nu} \quad \dots\dots\dots\dots(16).$$

If it be desired to emphasize the lower limit of summation we may use the notation ${}_\rho\mu_s$.

Differentiating the equation (16) with respect to q we obtain

$$\mu_s'=\sum_{\nu=\rho}^{r}\left\{\frac{(\nu-rq)^{s+1}}{qp}T_\nu-rs(\nu-rq)^{s-1}T_\nu\right\},$$

hence

$$\mu_{s+1}=qp\left[rs\mu_{s-1}+\mu_s'\right] \quad \dots\dots\dots\dots(17).$$

This I think is the simplest proof of which Romanovsky's formula is capable. It is a very interesting fact that the formula holds good without any alteration, even for the general case of incomplete moments.

For the incomplete moment of the first order $\mu_1=\sum_{\nu=\rho}^{r}(\nu-rq)T_\nu$ I have given the explicit expression in the *Skandinavisk Aktuarietidsskrift**,

$$\mu_1=\rho pT_\rho=\rho p\binom{r}{\rho}q^\rho p^{r-\rho} \quad \dots\dots\dots\dots(18),$$

Introducing this expression in the equation obtained from (17) by putting $s=0$, we get

$$\mu_0'=\frac{\rho}{q}T_\rho.$$

Integrating between 0 and q we obtain in the case $0<\rho$,

$$\mu_0=\rho\binom{r}{\rho}\int^q x^{\rho-1}(1-x)^{r-\rho}dx \quad \dots\dots\dots\dots(19).$$

This is the expression for the zero moment of the point binomial first found by Professor Pearson†. r and ρ being positive integers, the integral (19) may of course be calculated directly, but the expression for μ_0 obtained would not be any simpler than the definition

$$\mu_0=\sum_{\nu=\rho}^{r}T_\nu‡.$$

The formula (19) may still be useful in numerical applications as shown by Dr Camp§. Now assuming the numerical value of μ_0 to be calculated, I proceed to prove a recurrence formula by which the values of the higher incomplete moments may be calculated successively.

---

* No. 3, 1924, p. 161. † *Biometrika*, May 1924, Vol. XVI. p. 202, and note by Dr Camp, p. 171.

‡ The most interesting results of the direct integration of (19) are obtained by equating coefficients of equal powers of q. In this way some not unimportant identities satisfied by the binomial coefficients may be obtained.

§ *Biometrika*, May 1924, Vol. XVI. p. 157.

Putting

$$f_\nu = (\nu - rq)^{s-1},$$

$$g_\nu = (\nu - rq)\, T_\nu,$$

we have, by partial summation of (16),

$$\mu_s = \sum_{\nu=\rho}^{r} f_\nu g_\nu = f_r \sum_{\nu=\rho}^{r} g_\nu - \sum_{\nu=\rho}^{r-1} \Delta f_\nu \sum_{i=\rho}^{\nu} g_i .$$

Introducing the following expressions taken from (18),

$$\sum_{\nu=\rho}^{r} g_\nu = \rho p\, T_\rho ,$$

$$\sum_{i=\rho}^{\nu} g_i = \rho p\, T_\rho - (\nu+1)\, p T_{\nu+1},$$

we get

$$\mu_s = \rho p T_\rho f_r - \rho p\, T_\rho \sum_{\nu=\rho}^{r-1} \Delta f_\nu + \sum_{\nu=\rho}^{r-1} p\,(\nu+1)\, T_{\nu+1} \Delta f_\nu = \rho p\, T_\rho f_\rho + q \sum_{\nu=\rho}^{r} (r-\nu)\, T_\nu \Delta f_\nu$$

$$= \rho p\, T_\rho f_\rho + q \sum_{\nu=\rho}^{r} (rq - \nu + rp)\, T_\nu \Delta f_\nu = \rho p\, T_\rho f_\rho + rqp \sum_{\nu=\rho}^{r} T_\nu \Delta f_\nu - q \sum_{\nu=\rho}^{r} (\nu - rq)\, T_\nu \Delta f_\nu .$$

Introducing $\Delta f_\nu = \binom{s-1}{1} (\nu - rq)^{s-2} + \binom{s-1}{2} (\nu - rq)^{s-3} + \ldots + \binom{s-1}{s-1}$

we finally obtain

$$\mu_s = \rho p\, T_\rho (\rho - rq)^{s-1} + rqp \sum_{i=0}^{s-2} \binom{s-1}{i} \mu_i - q \sum_{i=0}^{s-2} \binom{s-1}{i} \mu_{i+1} \qquad \text{..............(20).}$$

This is Pearson's formula generalized to the case of incomplete moments. The formula only differs from the complete-moment formula (1) by the additional term $\rho p\, T_\rho (\rho - rq)^{s-1}$. Putting $\rho = 0$ in (20) we get of course (1). The incomplete moments do not satisfy any simple relation such as (5), so the equation (20) cannot be replaced by such simple reduced equations as (14*a*) and (14*b*).

The first few incomplete moments as calculated from (20) are

$$\mu_1 = \rho p\, T_\rho ,$$

$$\mu_2 = \rho p\, T_\rho [\rho - (r+1) q] + rqp \,.\, \mu_0 ,$$

$$\mu_3 = \rho p\, T_\rho [(\rho - (r+1) q)^2 + qp\,(2r-1)] + rqp\,(p-q) \,.\, \mu_0 ,$$

Review: The Elements of Vital Statistics in their bearing on Social and Public Health Problems. By Sir ARTHUR NEWSHOLME, K.C.B., M.D., F.R.C.P. New Edition, George Allen and Unwin, Ltd.

This new edition of *The Elements of Vital Statistics* contains much fresh matter. Sir Arthur Newsholme discusses in an early chapter the different methods of estimating population. We think that he has some doubt as to the advantage of substituting "age in years and months" for "age last birthday" in the 1921 census. Personally we doubt whether it will conduce to accuracy. If anyone knows his age "age last birthday" seems to us the simplest form in which to state that age and one wonders how recorders deal with a portion of a month when they have to enter "age in years and months." In considering "age at marriage" Sir Arthur Newsholme shows that between 1896 and 1920 there has been very little postponement of marriage, ·88 of a year for bachelor bridegrooms and ·40 of a year for spinster brides, and one concludes that such a slight postponement can have little to do with the falling birth-rate.

The birth-rate is considered in Chapters VII, VIII, and IX. In connection with the registration of births we should like to point out that though the maiden surname of the mother is required in Britain it is not always required when the birth of a British subject born abroad is registered and we think steps should be taken to remedy this defect in birth registration out of Britain where it occurs. We cannot see that there need be any connection between the sex ratios at birth and wholesale prices (p. 88) since both are varying with time. Later in the book there is a warning on this point but it is omitted in this instance.

As eugenists Sir Arthur Newsholme will not expect us to agree with much of what he writes on pp. 109, 110, and 111, though we fancy we see some "signs of grace" scattered here and there throughout the book and even on the pages referred to above! In the discussion on Infant Mortality it is assumed that a lower infant death-rate means improved health in the survivors; we still dare to think that this is the expression of a pious hope which is at present unproven. The treatment of life-tables is very superficial*.

We wish that heed would be taken of the chapter on the Registration of Sickness; we admit the difficulties but they should not be insuperable and the same seems to us to apply to the problem of death certificates though we think there will always be a difficulty over international statistics, even if we in England adopted the Swiss system. We notice on p. 290 that "Nothing is more certain than the statement that the removal of poverty would effect an enormous reduction of disease"; and in the next paragraph but one a quotation from the *Report of the Royal Commission on the Poor Laws* (1909) which states that "Sickness is therefore admittedly one of the chief causes of pauperism"; these two quotations seem to us in opposition and both require investigation before anyone can say which is the truer statement. Later on when it comes to the question of infant mortality Sir Arthur Newsholme points out the difficulty of disentangling cause and effect when poverty and sickness occur side by side; but though he has pointed this out he does not hesitate to state that "Environment has a preponderant influence in determining the magnitude of infantile and child mortality in a community"; by environment is meant, we presume, poverty, housing conditions, &c. Material collected by medical officers of health in individual districts seems to show that the most important factors are the health of the parents and the care that they bestow on their children, and though we consider it not yet proven, we are inclined to think that the fall in the infant death-rate of late years is due to the Midwives' Act and possibly, though hitherto there is no statistical proof, to some improvement in the habits of the mothers.

Again we cannot see that there need be any connection between prices and phthisis or that it is necessary to try to explain differences between England and Ireland in this connection. Sir Arthur Newsholme does not assume that there is a close connection between prices and tuberculosis but we think that to give coefficients of correlation when there is a time factor introduced is very dangerous unless they are by one method or another corrected for time. It would be interesting to know whether the number of people attacked by phthisis had diminished or not but such information would be, we think, impossible to obtain, since tuberculosis is now diagnosed at an earlier stage than formerly. The old question of the influence of heredity in the development of phthisis comes up again; some people among whom we count ourselves still believe in an hereditary predisposition but if one assumes that the association between tuberculosis in parent and child is due to large doses of infection at an early age rather than to an hereditary predisposition the argument against the phthisical having children seems to us equally strong; the children will be more likely to have tuberculosis than the children of parents without the disease.

In a review in *Biometrika* the statements in connection with the *Variate Difference Method* on p. 524 cannot be ignored. Sir Arthur Newsholme has quoted from a paper by Prof. Person of Harvard in which Prof. Person says that "the defective character of the variate difference method, as originally proposed, is admitted." We think that Sir Arthur Newsholme has not read the paper in *Biometrika* to which Prof. Person refers. There was an error in the application of the method to the problem of infant death-rate which the authors of that paper admitted but that is a very different matter from an error in the method itself. The product-moment method of obtaining a correlation coefficient is perfectly sound but it may be applied without taking into account the time factor or that regression is non-linear, and erroneous conclusions may be drawn from the coefficient of correlation found; such applications do not invalidate the method, and it has yet to be proved that the variate difference method is fundamentally at fault. We cannot help wishing that Sir Arthur Newsholme had ended his book at p. 500. There are many useful notes and "cautions" in the last 100 pages but also very much that seems to us open to criticism. We think the root of the difference between Sir Arthur Newsholme and the biometricians is to be found on p. 525 where it is suggested that "when the conclusions drawn from coefficients of correlation run counter to inferences on the same subject based on biological (including medical) considerations" judgment should be suspended. This seems to the statistician a most dangerous course to pursue. That a careful investigation of the results obtained by statistical methods is necessary we should all agree, but when this has been done we feel that the results must be accepted and that to accept only those results which agree with existing prejudgments is a retrograde step which must lead to disaster. We think it is this underlying idea which leads Sir Arthur Newsholme to reject evidence provided in the case of two collections of material and accept results deduced from another far more scanty and incomplete. On p. 558 the following sentence occurs: "Another example, on a somewhat scanty scale, may be quoted from McDougall's book. Miss Aylitt tested 342 children from the primary grades of schools in one district," but there is no reference to scantiness of material when the intelligence of 372 girls in one school and of 247 boys in another are tested and far-reaching conclusions are drawn by Dr Isserlis from this material. Sir Arthur Newsholme assumes that the teachers in the two schools referred to were really testing intelligence and not knowledge and one infers that he thinks that in all previous investigations it was the other way round! The correlations found with age make one doubtful and the comparatively low association between the teacher's estimate of intelligence and intelligence judged by a test makes one still more doubtful, but the numbers, reduced in these tables to 161 girls and 106 boys, are quite inadequate. We fear that it is because the conclusions reached do run counter to previous inferences on the subject that they are accepted uncritically.

---

* The result on p. 238 for p_x has been obtained from *English Life Tables*, 1914, p. 60, without Sir Arthur Newsholme troubling to test his own arithmetic. Had he done so, he would, perhaps, better have understood the meaning of the ordinary life-table symbols.

This book gives an excellent summary of what material is available in the census returns and shows how the statistics in this country could be improved in many ways, as for instance by the addition of the age of the two parents when the birth of an infant is registered, but we feel obliged to object to many of the conclusions, considering that some are contrary to existing evidence and that others are quite unproven. The book is bound to be replaced at no distant date, by one on sounder lines.

E. M. E.

Lebenserwartung und mittleres Alter der Lebenden.

VON E. I. GUMBEL, Heidelberg.

IM folgenden sollen die Grössenbeziehungen zwischen der zu einem bestimmten Alter gehörigen Lebenserwartung und dem mittleren Alter der nach der Sterbetafel über diesem Alter Lebenden untersucht werden. Für das Alter Null ist bereits in einer früheren Arbeit[*] gezeigt, dass für alle heutigen Sterbetafeln die Lebenserwartung eines Neugeborenen grösser ist, als das mittlere Alter aller Lebenden. Ein Ansatz zur Erweiterung dieses Satzes findet sich in einer Arbeit von Pearson[†], der gezeigt hat, dass bei der Gompertz-Makehamschen Absterbeordnung für jedes Alter die Summe Alter plus Lebenserwartung grösser ist als das hierzu gehörige mittlere Alter der Lebenden. Es soll nun untersucht werden, unter welchen allgemeinen Bedingungen dieser Satz gilt. Ferner soll auch eine untere Grenze für das mittlere Alter der Lebenden aufgestellt werden.

---

[*] "Uber eine Eigenschaft der Sterbetafel," *Zeitschrift für die gesamte Versicherungswissenschaft*, Bd. XXIV. H. 4, 1924.

[†] *Biometrika*, Vol. XVI. p. 297, 1924.

Man bezeichnet mit $l(x)$ die Wahrscheinlichkeit, die einem Neugeborenen zukommt, den $x$ten Geburtstag zu erreichen, sodass $l(0)=1$. Man definiere das höchste erreichbare Alter $\omega-1$ durch die zweite Grenzbedingung $l(\omega)=0$. Ferner definiere man die Sterbensintensität $\mu(x)$ wie üblich, als

$$\mu(x)=-\frac{d\,l(x)}{l(x)\,dx}.$$

Das mittlere Alter für die nach der Sterbetafel über dem Alter x Lebenden lautet

$$\overline{x(x)}=\frac{\int_x^\omega y\,l(y)\,dy}{\int_x^\omega l(y)\,dy}$$

Die Integrationsvariable wird hier stets mit y bezeichnet, um Verwechselungen mit der unteren Grenze x zu vermeiden. Die Lebenserwartung wird wie üblich definiert als

$$E(x)=\frac{1}{l(x)}\int_x^\omega l(y)\,dy.$$

Man bezeichne die Summe Alter + Lebenserwartung als Alterserwartung.

Somit gilt für die Gompertz-Makehamsche Absterbeordnung nach Pearson

$$\overline{x(x)}<x+E(x).$$

Dieser Satz lässt sich nach einem von Herrn v. Bortkiewicz* stammenden Verfahren für alle Absterbeordnungen beweisen, für welche die Sterbensintensität stets mit dem Alter wächst. Zum Beweise sei zunächst eine einfache Beziehung abgeleitet, die im folgenden von nöten ist. Man betrachte, nach der Definition der Sterbensintensität,

$$\int_x^\omega y^m\mu(y)\,l(y)\,dy=-\int_x^\omega y^m\,d\,l(y)$$

$$=x^m\,l(x)+m\int_x^\omega y^{m-1}\,l(y)\,dy.$$

Im Spezialfall

$m=0$: wird $\int_x^\omega \mu(y)\,l(y)\,dy=l(x)$,

$m=1$: wird $\int_x^\omega y\,\mu(y)\,l(y)\,dy=x\,l(x)+\int_x^\omega l(y)\,dy=l(x)\,(x+E(x))$,

$m=2$: wird $\int_x^\omega y^2\mu(y)\,l(y)\,dy=x^2\,l(x)+\frac{2\int_x^\omega y\,l(y)\,dy}{\int_x^\omega l(y)\,dy}\int_x^\omega l(y)\,dy=l(x)\,(x^2+2\overline{x(x)}\;E(x))$.

Unsere Aufgabe ist nun, Alterserwartung und mittleres Alter zu vergleichen. Man bezeichne

$$\begin{aligned}\sigma&=\operatorname{sign}\,[x+E(x)-\overline{x(x)}],\\&=\operatorname{sign}\,[l(x)\,(x+E(x))-l(x)\,\overline{x(x)}],\quad\text{da } l(x)>0,\\&=\operatorname{sign}\left[\int_x^\omega(y-\overline{x(x)})\,\mu(y)\,l(y)\,dy\right]\end{aligned}$$

nach den eben für $m=0$ und $m=1$ abgeleiteten Formeln.

Ferner ist: $$\sigma=\operatorname{sign}\left[\mu(\phi)\int_x^{\overline{x(x)}}(y-\overline{x(x)})\,l(y)\,dy+\int_{\overline{x(x)}}^\omega(y-\overline{x(x)})\,\mu(y)\,l(y)\,dy\right]$$

---

* L. v. Bortkiewicz, *Die mittlere Lebensdauer*, Jena, G. Fischer, 1893, p. 77.

nach dem Mittelwertsatz der Integralrechnung, wobei $x < \phi < \overline{x(x)}$. Aber nach der Definition des mittleren Alters der Lebenden ist

$$\int_x^{\overline{x(x)}} (y - \overline{x(x)})\, l(y)\, dy = -\int_{\overline{x(x)}}^{\omega} (y - \overline{x(x)})\, l(y)\, dy,$$

sodass

$$\begin{aligned} \sigma &= \operatorname{sign}\left[\int_{\overline{x(x)}}^{\omega} (y - \overline{x(x)})\,(\mu(y) - \mu(\phi))\, l(y)\, dy\right] \\ &= \operatorname{sign} \mu'(x) \end{aligned}$$

da wegen der Integrationsgrenzen $\phi < \overline{x(x)} < y$ und da $l(x)$ stets positiv.

Also gilt: *Die Alterserwartung ist grösser, gleich oder kleiner als das mittlere Alter der Lebenden, wenn die Sterbensintensität für alle Alter wächst, gleich bleibt oder fällt.* Dieser Satz lässt sich aber nicht umkehren. Es kann sehr wohl die Alterserwartung grösser sein als das mittlere Alter der Lebenden, ohne dass die Sterbensintensität stets wächst.

Der Pearsonsche Satz ergibt sich als Spezialfall. Denn die Gompertz-Makehamsche Formel geht aus von

$$\mu(x) = A + Bc^x,$$

woraus

$$\mu'(x) = Bc^x \log c < 0,$$

weil

$$B > 0,$$

und

$$c > 1.$$

Somit haben wir eine unter einer Bedingung gültige obere Grenze für das mittlere Alter der Lebenden aufgestellt. Nach einem allgemein für solche Abschätzungen verwandten Verfahren ist es leicht, auch eine untere Grenze hierfür aufzustellen. Man bilde die Wahrscheinlichkeit dafür, dass ein Neugeborener den $x$ten Geburtstag nicht erreicht,

$$g(x) = 1 - l(x),$$

dann gilt für die Momente dieser Funktion

$$\int_x^{\omega} y^m\, dg(y) = -\int_x^{\omega} y^m\, d\,l(y).$$

Man bilde nun den stets positiven Ausdruck $A = \frac{1}{l(x)} \int_x^{\omega} (a_1 + a_2 y)^2\, dg(y)$, wobei a_1 und a_2 vorläufig willkürliche reelle Konstanten sind. Führt man die Integration aus, so wird, wenn man die oben berechneten Momente verwendet,

$$A = a_1^2 + 2a_1 a_2 (x + E(x)) + a_2^2 (x^2 + 2\overline{x(x)}\, E(x)) > 0.$$

Nimmt man die quadratische Ergänzung vor, so wird

$$(a_1 + a_2 (x + E(x))^2 > a_2^2 (x + E(x))^2 - a_2^2 (x^2 + 2\overline{x(x)}\, E(x)).$$

Denkt man sich jedes einzelne Alter speziell betrachtet und setzt man hierfür jeweils

$$a_1 = -a_2 (x + E(x)),$$

so wird

$$x^2 + 2\overline{x(x)}\, E(x) > x^2 + 2x\, E(x) + E(x)^2, \text{ da } a_2^2 > 0,$$

oder

$$\overline{x(x)} > x + \frac{E(x)}{2}.$$

Demnach gilt der Satz: *Das mittlere Alter der nach der Sterbetafel über dem Alter x Lebenden ist grösser als das Alter x + der halben Lebenserwartung.* Für das Alter null gilt speziell: das mittlere Alter aller nach der Sterbetafel Lebenden ist grösser als die halbe Lebenserwartung eines Neugeborenen. Somit ist es gelungen, das mittlere Alter der Lebenden mit Hilfe der Lebenserwartung in feste Grenzen einzuschliessen.

FURTHER CONTRIBUTIONS TO THE THEORY OF SMALL SAMPLES*.

BY KARL PEARSON, F.R.S.

I. *On the Correlation of the two Standard Deviations and of either of these with the Coefficient of Correlation in small Samples taken from a bi-variate Population following the Normal Law of Distribution.*

THE distribution of the two standard deviations and the coefficient of correlation, namely σ_1, σ_2 and r, in samples of n taken from a normal population of constants Σ_1, Σ_2 and ρ is known to be†:

$$z = z_0 e^{\frac{1}{2(1-\rho^2)}\left(\frac{n\sigma_1^2}{\Sigma_1^2} - \frac{2nr\rho\sigma_1\sigma_2}{\Sigma_1\Sigma_2} + \frac{n\sigma_2^2}{\Sigma_2^2}\right)} \left(\frac{\sigma_1\sigma_2}{\Sigma_1\Sigma_2}\right)^{n-2} (1-r^2)^{\frac{n-4}{2}} \quad \ldots\ldots\ldots\text{(i)}.$$

Let us write

$$s_1 = \Sigma_1\sqrt{1-\rho^2}/\sqrt{n}, \quad s_2 = \Sigma_2\sqrt{1-\rho^2}/\sqrt{n}, \quad n\rho\sigma_1\sigma_2/\Sigma_1\Sigma_2(1-\rho^2) = \rho\sigma_1\sigma_2/s_1s_2;$$

then

$$z = z_0' e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}} e^{-\frac{1}{2}\frac{\sigma_2^2}{s_2^2}} \left(\frac{\sigma_1\sigma_2}{s_1s_2}\right)^{n-2} e^{\gamma r}(1-r^2)^{\frac{n-4}{2}} \quad \ldots\ldots\ldots\ldots\ldots\text{(ii)},$$

where z_0' differs from z_0 by terms not containing σ_1, σ_2 or r, and $\gamma = \rho\sigma_1\sigma_2/s_1s_2$.

Accordingly the frequency surface for σ_1, σ_2 will be obtained by integrating r throughout its entire range, i.e. from -1 to $+1$. We require accordingly the integral

$$I = \int_{-1}^{+1} e^{\gamma r}(1-r^2)^{\frac{n-4}{2}} dr \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)}$$

Take $r = \sin\theta$, and we have:

$$\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} e^{\gamma\sin\theta}\cos^{n-3}\theta\, d\theta.$$

Expand $e^{\gamma\sin\theta}$ and note that all terms

$$\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \sin^{2p+1}\theta\cos^{n-3}\theta\, d\theta = 0.$$

Accordingly

$$I = 2\int_0^{\frac{\pi}{2}}\left(1 + \frac{\gamma^2}{2!}\sin^2\theta + \frac{\gamma^4}{4!}\sin^4\theta + \ldots\right)\cos^{n-3}\theta\, d\theta.$$

---

* Printed from Lecture Notes.

† R. A. Fisher, *Biometrika*, Vol. x. p. 510, with a different notation.

Put $x = \sin^2\theta$ and we have

$$I = \int_0^1 \left(x^{-\frac{1}{2}} (1-x)^{\frac{n-4}{2}} + \frac{\gamma^2}{2!} x^{\frac{1}{2}} (1-x)^{\frac{n-4}{2}} + \frac{\gamma^4}{4!} x^{\frac{3}{2}} (1-x)^{\frac{n-4}{2}} + \dots \right.$$

$$\left. + \frac{\gamma^{2p}}{2p!} x^{\frac{2p-1}{2}} (1-x)^{\frac{n-4}{2}} + \dots \right) dx$$

$$= B\left(\frac{1}{2}, \frac{n-2}{2}\right) + \frac{\gamma^2}{2!} B\left(\frac{3}{2}, \frac{n-2}{2}\right) + \frac{\gamma^4}{4!} B\left(\frac{5}{2}, \frac{n-2}{2}\right) + \dots$$

$$+ \frac{\gamma^{2p}}{2p!} B\left(\frac{2p+1}{2}, \frac{n-2}{2}\right) + \dots \quad \dots\dots\text{(iv)},$$

where $B(p, q)$ is the usual beta-function,

$$= B\left(\frac{1}{2}, \frac{n-2}{2}\right) \left(1 + \frac{\gamma^2}{2!} \frac{1}{n-1} + \frac{\gamma^4}{4!} \frac{3.1}{(n-1)(n+1)} + \dots \right.$$

$$\left. + \frac{\gamma^{2p}}{2p!} \frac{(2p-1)(2p-3)\dots 1}{(n-1)(n+1)\dots(n+2p-3)} + \dots \right)$$

$$= B\left(\frac{1}{2}, \frac{n-2}{2}\right) \left(1 + \frac{\gamma^2}{1!} \cdot \frac{1}{2n-2} + \frac{\gamma^4}{2!} \cdot \frac{1}{(2n-2)(2n+2)} \right.$$

$$\left. + \frac{\gamma^6}{2!} \cdot \frac{1}{(2n-2)(2n+2)(2n+6)} + \dots + \frac{\gamma^{2p}}{p!} \cdot \frac{1}{(2n-2)(2n+2)\dots(2n+4p-6)} + \dots \right).$$

We can now return to (ii) and write down our equation to the frequency surface for σ_1, σ_2. It is

$$y = y_0 e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}} e^{-\frac{1}{2}\frac{\sigma_2^2}{s_2^2}} \left(\frac{\sigma_1\sigma_2}{s_1 s_2}\right)^{n-2} \left(1 + \frac{\rho^2}{1!} \frac{\sigma_1^2\sigma_2^2}{s_1^2 s_2^2} \frac{1}{2n-2} \right.$$

$$+ \frac{\rho^4}{2!} \frac{\sigma_1^4\sigma_2^4}{s_1^4 s_2^4} \frac{1}{(2n-2)(2n+2)} + \dots$$

$$\left. + \frac{\rho^{2p}}{p!} \frac{\sigma_1^{2p}\sigma_2^{2p}/s_1^{2p}s_2^{2p}}{(2n-2)(2n+2)\dots(2n+4p-6)} + \dots \right) \quad \dots\dots\dots\dots\dots\text{(v)}.$$

It is clear that in order to find any product moment function of this surface *about the* zero of σ_1, σ_2 we need only to find the value of

$$I' = \int_0^\infty e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}} \left(\frac{\sigma_1}{s_1}\right)^{2p} \left(\frac{\sigma_1}{s_1}\right)^m d\sigma_1$$

$$= s_1 \times \int_0^\infty e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}} \left(\frac{\sigma_1}{s_1}\right)^{2p+m} d\frac{\sigma_1}{s_1}.$$

Write $\qquad z = \frac{1}{2}\frac{\sigma_1^2}{s_1^2}$ or $\frac{\sigma_1}{s_1} = \sqrt{2z}.$

Hence I' is given by

$$I' = 2^{2p+m-1} s_1^{m+1} \int_0^\infty e^{-z} z^{\frac{2p+m-1}{2}} dz$$

$$= 2^{\frac{2p+m-1}{2}} s_1 \times \Gamma\left(\frac{2p+m+1}{2}\right) \quad \dots\dots\dots\dots\dots\dots\dots\text{(vi)}.$$

If N be the total number of samples, we have:

$$N=\int_0^\infty\int_0^\infty y\,d\sigma_1 d\sigma_2$$

$$=y_0 2^{n-3}(s_1 s_2)\times\left\{\Gamma^2\left(\frac{n-1}{2}\right)+\rho^2 2^2\,\Gamma^2\left(\frac{n+1}{2}\right)\frac{1}{1!\,(2n-2)}\right.$$

$$+\rho^4 2^4\,\Gamma^2\left(\frac{n+3}{2}\right)\frac{1}{2!\,(2n-2)(2n+2)}+\ldots$$

$$\left.+\rho^{2p}2^{2p}\,\Gamma^2\left(\frac{2p+m+1}{2}\right)\frac{1}{p!\,(2n-2)(2n+2)\ldots(2n+4p-6)}+\ldots\right\}$$

$$=y_0 2^{n-3}(s_1 s_2)\times\Gamma^2\left(\frac{n-1}{2}\right)\left(1+\frac{n-1}{2}\frac{\rho^2}{1!}+\frac{n-1}{2}\frac{n+1}{2}\frac{\rho^4}{2!}+\ldots\right.$$

$$\left.+\frac{n-1}{2}\frac{n+1}{2}\ldots\frac{n-1+2p}{2}\frac{\rho^{2p}}{p!}+\ldots\right).$$

Or
$$N=y_0\frac{2^{n-3}(s_1 s_2)\times\Gamma^2\left(\frac{n-1}{2}\right)}{(1-\rho^2)^{\frac{n-1}{2}}}\quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(vii)}.$$

This determines y_0, i.e.

$$y_0=\frac{N(1-\rho^2)^{\frac{n-1}{2}}}{2^{n-3}(s_1 s_2)\times\Gamma^2\left(\frac{n-1}{2}\right)}\quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(viii)}.$$

We shall now find

$$Np'_{\sigma_1\sigma_2}=\int_0^\infty\int_0^\infty y\sigma_1\sigma_2\,d\sigma_1 d\sigma_2$$

$$=s_1^2 s_2^2\int_0^\infty\int_0^\infty y\frac{\sigma_1}{s_1}\frac{\sigma_2}{s_2}\,d\frac{\sigma_1}{s_1}\,d\frac{\sigma_2}{s_2}$$

$$=y_0 s_1^2 s_2^2\, 2^{n-2}\left\{\Gamma^2\left(\frac{n}{2}\right)+\frac{\rho^2}{1!}2^2\Gamma^2\left(\frac{n+2}{2}\right)\frac{1}{2n-2}\right.$$

$$+\frac{\rho^4}{2!}2^4\,\Gamma^2\left(\frac{n+4}{2}\right)\frac{1}{(2n-2)(2n+2)}+\ldots$$

$$+\frac{\rho^{2p}}{p!}2^{2p}\Gamma^2\left(\frac{n+2p}{2}\right)(2n-2)\overline{(2n+2)\ldots(2n+4p-6)}$$

$$=y_0 s_1^2 s_2^2\, 2^{n-2}\Gamma^2\left(\frac{n}{2}\right)\left(1+\frac{\tfrac12 n\tfrac12 n}{\tfrac12(n-1)\,.\,1!}\rho^2\right.$$

$$+\frac{\tfrac12 n(\tfrac12 n+1)\tfrac12 n(\tfrac12 n+1)}{\tfrac12(n-1)\tfrac12(n+1)\,.\,2!}\rho^4+\ldots$$

$$\left.+\frac{\tfrac12 n(\tfrac12 n+1)\ldots\tfrac12(n+p)\tfrac12 n(\tfrac12 n+1)\ldots\tfrac12(n+p)}{\tfrac12(n-1)\tfrac12(n+1)\ldots\tfrac12(n+2p-3)\,.\,p!}\rho^{2p}+\ldots\right)$$

$$=y_0 s_1^2 s_2^2\, 2^{n-2}\Gamma^2\left(\frac{n}{2}\right)F(\tfrac12 n,\ \tfrac12 n,\ \tfrac12(n-1),\ \rho^2),$$

where $F(\alpha, \beta, \gamma, \rho^2)$ is the usual notation for the hypergeometrical series. But by Euler's Theorem:

$$F(\tfrac{1}{2}n, \tfrac{1}{2}n, \tfrac{1}{2}(n-1), \rho^2) = (1-\rho^2)^{-\frac{1}{2}(n+1)} F(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n-1), \rho^2).$$

Accordingly we have:

$$Np'_{\sigma_1\sigma_2} = y_0 s_1^2 s_2^2\, 2^{n-2}\Gamma^2\left(\frac{n}{2}\right)(1-\rho^2)^{-\frac{1}{2}(n+1)} \times H(\rho^2),$$

where

$$H(\rho^2) = \left(1 + \frac{\rho^2}{1\,!}\frac{1}{2n-2} + \frac{\rho^4}{2\,!}\frac{1}{(2n-2)(2n+2)} + \dots \right.$$
$$\left. + \frac{\rho^{2p}}{p!}\frac{1}{(2n-2)(2n+2)\dots(2n+4p-6)} + \dots\right) \dots\text{(ix)}.$$

Accordingly

$$p'_{\sigma_1\sigma_2} = \frac{2s_1s_2}{1-\rho^2}\frac{\Gamma^2\left(\frac{n}{2}\right)}{\Gamma^2\left(\frac{n-1}{2}\right)} H(\rho^2) \dots\dots\dots\dots\dots\dots\text{(x)}.$$

Returning to Equation (v) I now proceed to integrate it for one value σ_2 of the two variates only. We have

$$\int_0^\infty y\,d\sigma_2 = y_0 e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}}\left(\frac{\sigma_1}{s_1}\right)^{n-2} \times s_2$$
$$\times \int_0^\infty e^{-\frac{1}{2}\frac{\sigma_2^2}{s_2^2}}\left(\frac{\sigma_2}{s_2}\right)^{n-2}\sum_0^\infty \frac{\rho^{2p}}{p\,!}\frac{\sigma_1^{2p}\sigma_2^{2p}}{s_1^{2p}s_2^{2p}}\frac{1}{(2n-2)(2n+2)\dots(2n+4p-6)}\,d\left(\frac{\sigma_2}{s_2}\right)$$
$$= y_0 s_2 e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}}\left(\frac{\sigma_1}{s_1}\right)^{n-2} \times \sum_0^\infty \frac{\rho^{2p}}{p\,!}\left(\frac{\sigma_1}{s_1}\right)^{2p}$$
$$\times 2^{\frac{2p+n-3}{2}}\,\Gamma\left(\frac{n-1}{2}+p\right)\frac{1}{(2n-2)(2n+2)\dots(2n+4p-6)}$$
$$= y_0 s_2 2^{\frac{n-3}{2}}\,\Gamma\left(\frac{n-1}{2}\right) e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2}\left(\frac{\sigma_1}{s_1}\right)^{n-2}\left\{1 + \frac{\rho^2}{1!}\left(\frac{\sigma_1}{s_1}\right)^2 \frac{n-1}{2(n-1)}\right.$$
$$\left. + \frac{\rho^4}{2!}\left(\frac{\sigma_1}{s_1}\right)^2\frac{(n-1)(n-2)}{2^2(n-1)(n-2)} + \dots\right\}$$
$$= y_0 s_2 2^{\frac{n-3}{2}}\,\Gamma\left(\frac{n-1}{2}\right)\left(\frac{\sigma_1}{s_1}\right)^{n-2} e^{-\frac{1}{2}\frac{\sigma_1^2}{s_1^2}+\frac{1}{2}\rho^2\frac{\sigma_1^2}{s_1^2}}.$$

But $s_1 = \frac{\Sigma_1}{n}\sqrt{1-\rho^2}$; hence substituting we have for the frequency distribution:

$$y = \frac{y_0 s_2 2^{\frac{n-3}{2}}}{} \frac{n^{\frac{n-2}{2}}}{(1-\rho^2)^{\frac{n-2}{2}}} e^{-\frac{1}{2}\frac{n\sigma_1^2}{\Sigma_1^2}}\left(\frac{\sigma_1}{\Sigma_1}\right)^{n-2}\Gamma\left(\frac{n-1}{2}\right),$$

or finally substituting for y_0 from (viii):

$$y = \frac{Nn^{\frac{n-1}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)\Sigma_1} e^{-\frac{1}{2}\frac{n\sigma_1^2}{\Sigma_1^2}} \left(\frac{\sigma_1}{\Sigma_1}\right)^{n-2} \quad \text{..................(xi).}$$

This can be checked by integrating with regard to σ_1 and gives at once $\int_0^\infty y d\sigma_1 = N$, as it should. It is of course the usual form for the distribution of standard deviations in samples of n from a normal univariate population*. We now proceed to find

$$\int_0^\infty y\sigma_1 d\sigma_1 \text{ and } \int_0^\infty y\sigma_1^2 d\sigma_1$$

or as we may write them:

$$N_{\sigma_1}\mu_1' \text{ and } N_{\sigma_1}\mu_2'.$$

We have
$$N_{\sigma_1}\mu_1' = y_0\Sigma_1^2 \int_0^\infty e^{-\frac{1}{2}\frac{n\sigma_1^2}{\Sigma_1^2}} \left(\frac{\sigma_1}{\Sigma_1}\right)^{n-1} d\left(\frac{\sigma_1}{\Sigma_1}\right),$$

where y_0 is written for brevity for

$$Nn^{\frac{n-1}{2}} \Big/ 2^{\frac{n-3}{2}} \Gamma\left(\frac{n-1}{2}\right) \Sigma_1.$$

Hence
$${}_{\sigma_1}\mu_1' = \frac{y_0}{N} \frac{\Sigma_1^2 2^{\frac{n-2}{2}}}{n^{\frac{1}{2}n}} \Gamma\left(\frac{n}{2}\right),$$

or substituting for y_0

$${}_{\sigma_1}\mu_1' = \Sigma_1 \sqrt{\frac{2}{n}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \quad \text{..........................(xii)}$$

Again
$$N_{\sigma_1}\mu_2' = y_0\Sigma_1^3 \int_0^\infty e^{-\frac{1}{2}\frac{n\sigma_1^2}{\Sigma_1^2}} \left(\frac{\sigma_1}{\Sigma_1}\right)^{n} d\left(\frac{\sigma_1}{\Sigma_1}\right)$$

$$= \frac{y_0\Sigma_1^3 2^{\frac{n-1}{2}}}{n^{\frac{n-1}{2}}} \Gamma\left(\frac{n+1}{2}\right),$$

or, on substituting for y_0, we find the familiar result

$${}_{\sigma_1}\mu_2' = \Sigma_1^2 . \frac{n-1}{n} \quad \text{..............................(xiii).}$$

Hence writing $\lambda_n = \Gamma\left(\frac{n}{2}\right) \Big/ \Gamma\left(\frac{n-1}{2}\right)$, we have

$${}_{\sigma_1}\mu_2 = \sigma^2_{\sigma_1} = \frac{\Sigma_1^2}{n}\left(n - 1 - 2\lambda_n^2\right) \quad \text{.......................(xiv).}$$

---

* *Biometrika*, Vol. VI. p. 304 and Vol. X. p. 522.

Accordingly

$$\left.\begin{aligned}\bar{x}_{\sigma_1}\bar{x}_{\sigma_2} &= \Sigma_1\Sigma_2\frac{2}{n}\lambda_n^2\\ \sigma_{\sigma_1}\sigma_{\sigma_2} &= \Sigma_1\Sigma_2\frac{1}{n}(n-1-2\lambda_n^2)\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots\text{(xv)}.$$

We can write (x) in the form:

$$p'_{\sigma_1\sigma_2} = \Sigma_1\Sigma_2\frac{2}{n}\lambda_n^2 H(\rho^2),$$

and as the correlation coefficient:

$$r_{\sigma_1\sigma_2} = \frac{p'_{\sigma_1\sigma_2} - \bar{x}_{\sigma_1}\bar{x}_{\sigma_2}}{\sigma_{\sigma_1}\sigma_{\sigma_2}},$$

we find

$$r = \frac{2\lambda_n^2(H(\rho^2)-1)}{n-1-2\lambda_n^2}\quad\ldots\ldots\ldots\ldots\text{(xvi)}.$$

We shall now proceed to investigate the value of the factor $2\lambda_n^2/(n-1-2\lambda_n^2)$ when $n-1$ is *large*. We shall have to be cautious because if $2\lambda_n^2$ starts with the term $(n-1)$ everything will depend on the *second* term in the expansion. We will therefore expand λ_n in terms of $\epsilon = \frac{1}{2}(n-1)$.

A good formula for $\Gamma(x)$ is that provided by Schlömilch*, namely:

$$\log\Gamma(x) = \tfrac{1}{2}\log 2\pi + (x-\tfrac{1}{2})\log x - x + \frac{1}{12}\frac{1}{x} - \frac{1}{360}\frac{1}{x(x+1)(x+2)} + \text{etc.}\quad\text{(xvii)}.$$

Now

$$\begin{aligned}\log\lambda_n &= \log\Gamma\left(\frac{n}{2}\right) - \log\Gamma\left(\frac{n-1}{2}\right)\\ &= \log\Gamma\left(\frac{n-1}{2}+\tfrac{1}{2}\right) - \log\Gamma\left(\frac{n-1}{2}\right)\\ &= \log\Gamma(\epsilon+\tfrac{1}{2}) - \log\Gamma(\epsilon),\end{aligned}$$

and this equals, by the above formula:

$$\begin{aligned}&\epsilon\log(\epsilon+\tfrac{1}{2}) - \epsilon - \tfrac{1}{2} + \frac{1}{12}\frac{1}{\epsilon+\frac{1}{2}} + \text{etc.}\\ &-(\epsilon-\tfrac{1}{2})\log\epsilon + \epsilon - \frac{1}{12}\frac{1}{\epsilon} + \text{etc.}\end{aligned}$$

or

$$\begin{aligned}\log\lambda_n &= \tfrac{1}{2}\log\epsilon + \epsilon\log\left(1+\frac{1}{2\epsilon}\right) - \tfrac{1}{2} + \text{terms in }\frac{1}{\epsilon^2}\text{ etc.}\\ &= \tfrac{1}{2}\log\epsilon + \tfrac{1}{2} - \frac{1}{8\epsilon} - \tfrac{1}{2} + \text{terms in }\frac{1}{\epsilon^2}\text{ etc.}\end{aligned}$$

Hence

$$\begin{aligned}\lambda_n &= \sqrt{\epsilon}\,e^{-\frac{1}{8\epsilon}} = \sqrt{\epsilon}\left(1-\frac{1}{8\epsilon}\right) + \text{etc.}\\ &= \frac{\sqrt{\epsilon}}{\left(1+\frac{1}{4\epsilon}\right)^{\frac{1}{2}}}\end{aligned}$$

to our degree of approximation.

---

* *Compendium der höheren Analysis*, Bd. II. S. 264.

Hence $$\lambda_n^2 \quad \frac{4\epsilon^2}{1+4\epsilon},$$

and $$\frac{2\lambda_n^2}{2\epsilon-2\lambda_n^2}=\frac{\lambda_n^2}{\epsilon-\lambda_n^2}=\frac{4\epsilon^2}{\epsilon}=2(n-1)+\text{etc.}$$

Accordingly when n is large:

$$r_{\sigma_1\sigma_2} \text{ approaches } (2n-2)(H(\rho^2)-1)$$
$$=\rho^2+\text{terms in inverse powers of } 2n+2,\ 2n+4, \text{ etc.}$$

This approaches the value ρ^2 given long ago* for *large* samples from material with linear regression and homoscedasticity.

In the ordinary theory of probable errors it is usually the mean value of the product $p_{\sigma_1\sigma_2}$ that is involved, i.e. $\sigma_{\sigma_1}\sigma_{\sigma_2}r_{\sigma_1\sigma_2}$. But σ_{σ_1}, σ_{σ_2} are known and have been tabled for small samples. It remains to table $r_{\sigma_1\sigma_2}$ and see how far it diverges from the value usually assumed (i.e. ρ^2) for it.

The following table gives the values of $r_{\sigma_1\sigma_2}$ in samples of size 2, 3, ... 25, 50, 100, 400 for correlations ρ in the sampled population of ·0, ·1, ·2, ... ·8, ·9.

TABLE I. *Values of the Correlation of Standard Deviations in Samples for Different Values of Variate Correlation in a Sampled Normal Population.*

Values of ρ.

Size of Sample.	*·0*	*·1*	*·2*	*·3*	*·4*	*·5*	*·6*	*·7*	*·8*	*·9*	*1·0*
2	·0000	·0088	·0352	·0794	·1420	·2236	·3249	·4470	·5912	·7587	·9514
3	·0000	·0092	·0367	·0828	·1484	·2323	·3369	·4622	·6094	·7795^{-}	·9737
4	·0000	·0094	·0375^{+}	·0846	·1509	·2369	·3430	·4699	·6184	·7893	·9837
5	·0000	·0095^{-}	·0380	·0857	·1528	·2396	·3467	·4749	·6235^{+}	·7948	·9890
6	·0000	·0096	·0383	·0864	·1540	·2415^{-}	·3491	·4775^{+}	.6268	·7981	·9921
7	·0000	·0096	·0386	·0870	·1549	·2427	·3508	·4794	·6291	·8004	·9940
8	·0000	·0097	·0388	·0873	·1556	·2437	·3520	·4808	·6307	·8020	·9953
9	·0000	·0097	·0389	·0876	·1561	·2444	·3530	·4820	·6319	·8031	·9962
10	·0000	·0097	·0390	·0879	·1565^{-}	·2450	·3537	·4829	·6328	·8040	·9969
11	·0000	·0098	·0391	·0881	·1568	·2455^{-}	·3543	·4836	·6336	·8047	·9975^{-}
12	·0000	·0098	·0392	·0882	·1571	·2459	·3548	·4842	·6342	·8053	·9978
13	·0000	·0098	·0392	·0884	·1573	·2462	·3552	·4847	·6347	·8058	·9982
14	·0000	·0098	·0393	·0885	·1575^{+}	·2465^{-}	·3556	·4851	·6351	·8061	·9984
15	·0000	·0098	·0393	·0886	·1577	·2467	·3559	·4854	·6355^{+}	·8065^{-}	·9986
16	·0000	·0098	·0394	·0887	·1578	·2469	·3562	·4857	·6358	·8067	·9988
17	·0000	·0099	·0394	·0888	·1580	·2471	·3564	·4860	·6361	·8070	·9989
18	·0000	·0099	·0395^{-}	·0888	·1581	·2473	·3566	·4862	·6366	·8072	·9990
19	·0000	·0099	·0395^{-}	·0889	·1582	·2474	·3568	·4864	·6363	·8074	·9991
20	·0000	·0099	·0395^{+}	·0890	·1583	·2476	·3570	·4866	·6367	·8075^{+}	·9992
21	·0000	·0099	·0395^{+}	·0890	·1584	·2477	·3571	·4868	·6369	·8077	·9993
22	·0000	·0099	·0396	·0891	·1584	·2478	·3573	·4870	·6371	·8078	·9993
23	·0000	·0099	·0396	·0891	·1585^{+}	·2479	·3574	·4871	·6372	·8079	·9994
24	·0000	·0099	·0396	·0891	·1586	·2480	·3575^{-}	·4872	·6373	·8080	·9994
25	·0000	·0099	·0396	·0892	·1586	·2481	·3576	·4873	·6375^{-}	·8081	·9995
50	·0000	·0100	·0398	·0896	·1593	·2490	·3588	·4887	·6388	·8091	·9999
100	·0000	·0100	·0399	·0898	·1597	·2495^{+}	·3594	·4894	·6394	·8095^{+}	·99997
400	·0000	·0100	·0400	·0900	·1599	·2499	·3599	·4898	·6399	·8099	1·00000
ρ^2	·0000	·0100	·0400	·0900	·1600	·2500	·3600	·4900	·6400	·8100	1·0000

---

* *Biometrika*, Vol. IX. p. 4. See also *Phil. Trans.* Vol. 191, A (1898), p. 242.

II. *On the Correlation of Deviations in the Standard Deviation of a Variate and the Correlation Coefficient of Variates in Samples from a Normal Population.*

We start from the general surface for the three variates, i.e.:

$$z = z_0' e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2} e^{-\frac{1}{2}\left(\frac{\sigma_2}{s_2}\right)^2} e^{\gamma r}(1-r^2)^{\frac{n-4}{2}} \left(\frac{\sigma_1}{s_1}\right)^{n-2} \left(\frac{\sigma_2}{s_2}\right)^{n-2},$$

and integrate it for the variate σ_2 between the limits 0 and ∞. We shall thus reach the bivariate frequency surface for σ_1 and r. In order to do this we must expand the exponential $e^{\gamma r}$ and find:

$$z' = \int_0^\infty z\, d\sigma_2 = z_0' e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2} \left(\frac{\sigma_1}{s_1}\right)^{n-2} (1-r^2)^{\frac{n-4}{2}} \int_0^\infty e^{-\frac{1}{2}\left(\frac{\sigma_2}{s_2}\right)^2} \left(\frac{\sigma_2}{s_2}\right)^{n-2} \overset{\infty}{\underset{0}{S}} \frac{(r\rho)^p}{p!} \left(\frac{\sigma_1\sigma_2}{s_1 s_2}\right)^p d\sigma_2.$$

Remembering that the integral

$$\int_0^\infty e^{-\frac{1}{2}\left(\frac{\sigma_2}{s_2}\right)^2} \left(\frac{\sigma_2}{s_2}\right)^{n-2+p} d\sigma_2 = s_2 2^{\frac{n-3+p}{2}} \Gamma\left(\frac{n-1+p}{2}\right),$$

we can write for the frequency surface of σ_1 and r

$$z' = z_0' s_2 e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2} \left(\frac{\sigma_1}{s_1}\right)^{n-2} (1-r^2)^{\frac{n-4}{2}} 2^{\frac{n-3}{2}} \left[\Gamma\left(\frac{n-1}{2}\right)\right.$$

$$\times \left\{1 + 2\,\frac{n-1}{2}\,\frac{(r\rho)^2}{2!}\left(\frac{\sigma_1}{s_1}\right)^2 + 2^2\,\frac{n-1}{2}\,\frac{n+1}{2}\,\frac{(r\rho)^4}{4!}\left(\frac{\sigma_1}{s_1}\right)^4 + \dots\right\}$$

$$\left. + \Gamma\left(\frac{n}{2}\right)\sqrt{2}\, r\rho\,\frac{\sigma_1}{s_1}\left\{1 + 2\,\frac{n}{2}\,\frac{(r\rho)^2}{3!}\left(\frac{\sigma_1}{s_1}\right)^2 + 2^2\,\frac{n}{2}\left(\frac{n}{2}+1\right)\frac{(r\rho)^4}{5!}\left(\frac{\sigma_1}{s_1}\right)^4 + \dots\right\}\right] \quad \text{...(xviii)}.$$

The complexity of this surface for r and σ_1 is at once obvious, and it is not easy to form a mental picture of the form of the surface for small values of n. On the other hand we shall find it relatively easy to evaluate the integrals

$$I_{00} = \int_{-1}^{+1}\int_0^\infty z'\,dr\,d\sigma_1, \quad I_{10} = \int_{-1}^{+1}\int_0^\infty z' r\,dr\,d\sigma_1, \quad I_{01} = \int_{-1}^{+1}\int_0^\infty z'\sigma_1\,dr\,d\sigma_1,$$

$$I_{20} = \int_{-1}^{+1}\int_0^\infty z' r^2\,dr\,d\sigma_1, \text{ and } I_{11} = \int_{-1}^{+1}\int_0^\infty z' r\sigma_1\,dr\,d\sigma_1.$$

The first four will check results already known, and the last will provide us with what we are at present seeking. When we integrate with regard to r, all the terms in odd powers of r disappear and we may take for the even powers the double integral from 0 to + 1. Remembering that

$$2\int_0^1 (1-r^2)^{\frac{n-4}{2}} r^{2p}\,dr = B\left(\frac{2p+1}{2}, \frac{n-2}{2}\right)$$

$$= \Gamma\left(\frac{2p+1}{2}\right)\Gamma\left(\frac{n-2}{2}\right)\Big/\Gamma\left(\frac{2p+n-1}{2}\right)$$

we easily find

$$\int_{-1}^{+1} z'dr = z_0's_2 e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2} \left(\frac{\sigma_1}{s_1}\right)^{n-2} 2^{\frac{n-3}{2}} \sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)$$

$$\times\left\{1+\frac{\rho^2}{2}\left(\frac{\sigma_1}{s_1}\right)^2\frac{1}{1!}+\frac{\rho^4}{2^2}\left(\frac{\sigma_1}{s_1}\right)^4\frac{1}{2!}+\dots\right\}$$

$$= z_0's_2 e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2} e^{+\frac{1}{2}\rho^2\frac{\sigma_1^2}{s_1^2}} 2^{\frac{n-3}{2}} \sqrt{\pi}\,\Gamma\left(\frac{n-2}{2}\right)\left(\frac{\sigma_1}{s_1}\right)^{n-2}$$

$$= z_0's_2 e^{-\frac{1}{2}(1-\rho^2)\frac{\sigma_1^2}{s_1^2}} \left(\frac{\sigma_1}{s_1}\right)^{n-2} \sqrt{\pi}\, 2^{\frac{n-3}{2}}\,\Gamma\left(\frac{n-2}{2}\right) \quad\dots\dots\dots\dots\dots\text{(xix)}.$$

If we integrate next with regard to σ_1 we have to find

$$\int_0^\infty e^{-\frac{1}{2}(1-\rho^2)\frac{\sigma_1^2}{s_1^2}}\left(\frac{\sigma_1}{s_1}\right)^{n-2} d\sigma_1,$$

which equals

$$s_1 2^{\frac{n-3}{2}}\frac{1}{(1-\rho^2)^{\frac{n-1}{2}}}\Gamma\left(\frac{n-1}{2}\right).$$

Accordingly we have

$$N = z_0's_1s_2\sqrt{\pi}\,2^{n-3}\,\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\Big/(1-\rho^2)^{\frac{n-1}{2}} \quad\dots\dots\dots\text{(xx)},$$

which checks with earlier results.

We will now investigate $\int_{-1}^{+1} z'r\,dr$, which throws us onto our second series in (xviii) for z':

$$\int_{-1}^{+1} z'r\,dr = z_0's_2 e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2}\left(\frac{\sigma_1}{s_1}\right)^{n-1} 2^{\frac{n-2}{2}}\rho\Gamma\left(\frac{n}{2}\right)\tfrac{1}{2}\sqrt{\pi}\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}$$

$$\times\left\{1+\frac{n}{n+1}\frac{\rho^2}{1!\,2}\left(\frac{\sigma_1}{s_1}\right)^2+\frac{n(n+2)}{(n+1)(n+3)}\frac{\rho^4}{2!\,2^2}\left(\frac{\sigma_1}{s_1}\right)^4+\dots\right\} \quad\dots\text{(xxi)}.$$

If we divide this result by that in (xx) we obtain the mean value of r for a given array of r's corresponding to σ_1, i.e.

$$\bar{r}_{\sigma_1} = \frac{\rho}{\sqrt{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\frac{\sigma_1}{s_1}e^{-\frac{1}{2}\rho^2\left(\frac{\sigma_1}{s_1}\right)^2}\left\{1+\frac{n}{n+1}\frac{\rho^2}{1!\,2}\left(\frac{\sigma_1}{s_1}\right)^2\right.$$

$$\left.+\frac{n(n+2)}{(n+1)(n+3)}\frac{\rho^4}{2!\,2^2}\left(\frac{\sigma_1}{s_1}\right)^4+\dots\right\} \quad\dots\dots\dots\dots\text{(xxii)}.$$

The regression line is therefore non-linear. Let us see what form it takes as we increase n. Since $s_1 = \Sigma_1\sqrt{1-\rho^2}/\sqrt{n}$, we must approach the limit with caution for s_1 becomes zero as n increases. Let us write $\kappa = \frac{1}{2}\rho^2\sigma_1^2/((1-\rho^2)\,\Sigma_1^2)$ which will be finite. Then we have to find the limit of

$$r_{\sigma_1} = \rho\,\frac{\sqrt{\frac{1}{2}n}\,\Gamma\left(\frac{n}{2}\right)}{\Gamma(\frac{1}{2}(n+1))}\,\frac{}{\Sigma_1\sqrt{1-\rho^2}}\,e^{-n\kappa}\left(1+\frac{n}{n+1}\,\frac{n\kappa}{1\,!}+\frac{n\,(n+2)}{(n+1)(n+3)}\,\frac{n^2\kappa^2}{2\,!}\right.$$
$$\left.+\frac{n(n+2)\,(n+4)}{(n+1)\,(n+3)\,(n+5)}\,\frac{n^3\kappa^3}{3\,!}+\dots\right).$$

But $\sqrt{\frac{1}{2}n}\,\Gamma\,(\frac{1}{2}n)/\Gamma\,(\frac{1}{2}\,(n+1))$ takes the limit unity.

Writing $n\kappa = \kappa'$, we have now to find the limit of:

$$= e^{-\kappa'}\left(1+\frac{n}{n+1}\,\frac{\kappa'}{1\,!}+\frac{n\,(n+2)}{(n+1)(n+3)}\,\frac{\kappa'^2}{2\,!}+\frac{n\,(n+2)\,(n+4)}{(n+1)\,(n+3)\,(n+5)}\,\frac{\kappa'^3}{3\,!}+\right.$$

Expanding the exponential and multiplying out we have:

$$L = 1-\frac{1}{n+1}\,\frac{\kappa'}{1\,!}+\frac{1\,.\,3}{(n+1)\,(n+3)}\,\frac{\kappa'^2}{2\,!}-\frac{1\,.\,3\,.\,5}{(n+1)\,(n+3)\,(n+5)}\,\frac{\kappa'^3}{3\,!}+\dots$$
$$= 1-\frac{n}{n+1}\,\frac{\kappa}{1\,!}+\frac{n^2\,1\,.\,3}{(n+1)\,(n+3)}\,\frac{\kappa^2}{2!}-\frac{n^3\,1\,.\,3\,.\,5}{(n+1)\,(n+3)\,(n+5)}\,\frac{\kappa^3}{3!}+\dots.$$

Hence proceeding to the limit we find:

$$L = 1-\frac{\kappa}{1\,!}+\frac{1\,.\,3}{2\,!}\,\kappa^2-\frac{1\,.\,3\,.\,5}{3\,!}\,\kappa^3+\dots$$
$$(1+2\kappa)^{-\frac{1}{2}} = \left(1+\frac{\rho^2}{1-\rho^2}\,\frac{\sigma_1^2}{\Sigma_1^2}\right)^{-\frac{1}{2}}$$

Or as n grows large we have nearly and more nearly:

$$\bar{r}_{\sigma_1} = \rho\,\frac{\sigma_1}{\Sigma_1}\,\frac{1}{\left(1-\rho^2+\rho^2\left(\frac{\sigma_1}{\Sigma_1}\right)^2\right)^{\frac{1}{2}}} \quad \text{.................(xxii } \textit{bis}\text{).}$$

Thus the equation suggested by normal correlation:

$$\bar{r}_{\sigma_1}-\bar{r} = \frac{\rho}{\sqrt{2}}\,\frac{\sigma_r}{\sigma_\sigma}\,(\sigma_1-\bar{\sigma}_1) = \rho\,.\,(1-\rho^2)\,\frac{\sigma_1-\bar{\sigma}_1}{\Sigma_1}\,,$$

i.e. the usual value for linear regression, is seen to be incorrect. The true regression is not linear, even for n indefinitely large. When $\rho = 1$, $\bar{r}_{\sigma_1} = 1$, from (xxii *bis*), as it should. This result seems of considerable importance. For, while the distribution of σ_1 approaches closely to a normal curve when n is large*, and the values of β_1, β_2 for the distribution of r approach the values 0 and 3, i.e. the normal values, for moderate values of ρ and values over 50 of n, they do not do so for ρ above 0·9. In other words the limit for n large in the case of (xviii) is not a normal surface. The appearance of such a regression curve as (xxii *bis*) seems to me of the highest importance, and provides a warning against the too ready use of linear regression for frequency constants.

---

* *Biometrika*, Vol. x. p. 524.

If we now integrate the expression in (xxi) for σ_1 we obtain

$$N\bar{r} = z_0 s_1 s_2 2^{n-3} \rho \sqrt{\pi} \frac{\Gamma^2\left(\frac{n}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}$$

$$\times\left(1 + \frac{\frac{1}{2}n \cdot \frac{1}{2}n}{\frac{1}{2}(n+1)}\frac{\rho^2}{1!} + \frac{\frac{1}{2}n\left(\frac{1}{2}n+1\right)\frac{1}{2}n\left(\frac{1}{2}n+1\right)}{\frac{1}{2}(n+1)\frac{1}{2}(n+3)}\frac{\rho^4}{2!} + \dots\right).$$

But the series $= F\left(\frac{1}{2}n, \frac{1}{2}n, \frac{1}{2}(n+1), \rho^2\right)$,

where $F(\alpha, \beta, \gamma, x)$ is the usual hypergeometrical series. Using Euler's transformation, we find:

$$N\bar{r} = \frac{z_0 s_1 s_2 2^{n-3} \rho \sqrt{\pi}}{(1-\rho^2)^{\frac{n-1}{2}}} \frac{\Gamma^2\left(\frac{n}{2}\right)\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} F\left(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n+1), \rho^2\right).$$

Whence dividing by N from (xx)

$$\bar{r} = \rho \frac{\Gamma^2\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{n+1}{2}\right)} F\left(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n+1), \rho^2\right) \text{.............(xxiii).}$$

This agrees with *Biometrika*, Vol. XI. p. 336, Equation (xxv), if we change the notation and observe that the symbol q_n there used is the same as $B\left(\frac{1}{2}, \frac{1}{2}n\right)$.

It is extremely easy in the same manner to find $N\bar{\sigma}_1$. All we have to do is to integrate the expression in (xix) for $\sigma_1 d\sigma_1$, i.e.

$$N\bar{\sigma}_1 = \int_{-1}^{+1}\int_0^{\infty} z'\sigma_1 dr\, d\sigma_1 = z_0 s_1^2 s_2 \sqrt{\pi}\, 2^{\frac{n-3}{2}} \Gamma\left(\frac{n-2}{2}\right) \times \int_0^{\infty} e^{-\frac{1}{2}(1-\rho^2)\left(\frac{\sigma_1}{s_1}\right)^2} \left(\frac{\sigma_1}{s_1}\right)^{n-1} d\left(\frac{\sigma_1}{s_1}\right)$$

$$= \frac{z_0 s_1^2 s_2 \sqrt{\pi}\, 2^{\frac{2n-5}{2}}}{(1-\rho^2)^{\frac{1}{2}n}} \Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n}{2}\right).$$

Whence getting rid of z_0 we find:

$$\bar{\sigma}_1 = \frac{s_1\sqrt{2}}{\sqrt{1-\rho^2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = \Sigma_1\sqrt{\frac{2}{n}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \text{ ....(xxiv).}$$

This agrees with the result (xii) given on p. 180.

We now proceed to find:

$$Np'_{r\sigma_1} = \int_0^{\infty} \sigma_1 d\sigma_1 \int_{-1}^{+1} z' r\, dr.$$

We return to Equation (xxi) and integrate with regard to $\sigma_1 d\sigma_1$:

$$Np'_{r\sigma_1} = z_0 s_2 s_1^2 2^{\frac{n-4}{2}} \rho \sqrt{\pi} \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \Gamma\left(\frac{n}{2}\right) \times \int_0^{\infty} \left\{1 + \frac{n}{n+1}\frac{\rho^2}{1!\,2}\left(\frac{\sigma_1}{s_1}\right)^2\right.$$

$$\cdot \frac{n\,(n+2)}{(n+1)\,(n+3)}\,\frac{\rho^4}{2!\,2^2}\left(\frac{\sigma_1}{s_1}\right)^4 + \ldots\Bigg\}\, e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2}\left(\frac{\sigma_1}{s_1}\right)^n d\left(\frac{\sigma_1}{s_1}\right)$$

$$= z_0 s_2 s_1^2\, 2^{\frac{2n-5}{2}}\,\Gamma\left(\frac{n+1}{2}\right)\rho\sqrt{\pi}\,\frac{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}$$

$$\times\left(1 + 2\,\frac{n+1}{2}\,\frac{n}{n+1}\,\frac{\rho^2}{1!\,2} + 2^2\,\frac{n+1}{2}\,\frac{n+3}{2}\,\frac{n\,(n+2)}{(n+1)\,(n+3)}\,\frac{\rho^4}{2!\,2^2} + \ldots\right).$$

Or:
$$Np'_{r\sigma_1} = \frac{z_0 s_2 s_1^2\,\rho\sqrt{\pi}\,2^{\frac{}{2}}\,\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n}{2}\right.}{(1-\rho^2)^{\frac{n}{2}}}$$

Hence getting rid of z_0, we find

$$p'_{r\sigma_1} = \frac{s_1\rho\sqrt{2}}{\sqrt{1-\rho^2}}\,\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = \Sigma_1\rho\sqrt{\frac{2}{n}}\,\frac{\Gamma\left(\frac{2}{n}\right)}{} \qquad \ldots\ldots\ldots\ldots \text{(xxv)}.$$

We have already given in this paper the value of $\sigma_{\sigma_1} = \Sigma_1 (n - 1 - 2\lambda_n^2)^{\frac{1}{2}}/\sqrt{n}$, see Equation (xiv). In *Biometrika*, Vol. XI. p. 335, the value of σ_r^2 is given, but it may be well to deduce it by the methods here used. Clearly we must integrate z' for r^2dr, or better for $(1-(1-r^2))\,dr$ and then for $d\sigma_1$. It is seen at once that only the first series on the right of Equation (xviii) concerns us for the second contains odd powers. Accordingly we have

$$N(\sigma_r^2 + \bar{r}^2) = \int_{-1}^{+1} z'(1-(1-r^2))\,dr\,d\sigma_1$$

$$= N - 2z_0 s_2 \int_0^1\int_0^\infty (1-r^2)^{\frac{n-2}{2}}\,2^{\frac{n-3}{2}}\,\Gamma\left(\frac{n-1}{2}\right) e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2}\left(\frac{\sigma_1}{s_1}\right)^{n-2}$$

$$\times\ 1 + \frac{(n-1)}{2!}\,r^2\rho^2\left(\frac{\sigma_1}{s_1}\right)^2 + \frac{(n-1)(n+1)}{4!}\,r^4\rho^4\left(\frac{\sigma_1}{s_1}\right)^4 + \ldots\Bigg\}\,dr\,d\sigma_1.$$

We have now

$$2\int_0^1 (1-r^2)^{\frac{n-2}{2}}\,r^{2p}dr = \Gamma\left(\frac{2p+1}{2}\right)\Gamma\left(\frac{n}{2}\right)\Big/\,\Gamma\left(\frac{2p+n+1}{2}\right),$$

and accordingly

$$N(\sigma_r^2 + \bar{r}^2) = N - z_0 s_2\, 2^{\frac{n-3}{2}}\,.\,\frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}}{\Gamma\left(\frac{n+1}{2}\right)}\int_0^\infty e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2}\left(\frac{\sigma_1}{s_1}\right)$$

$$\times\left\{1 + \tfrac{1}{2}\,\frac{n-1}{2!}\,\rho^2\left(\frac{\sigma_1}{s_1}\right)^2\frac{1}{n+1} + \tfrac{1}{2}\cdot\tfrac{3}{2}\,\frac{(n-1)(n+1)}{4!}\,\rho^4\left(\frac{\sigma_1}{s_1}\right)^4\frac{1}{n+1\;n+3} + \ldots\right\}d\sigma_1.$$

We can now integrate with regard to σ_1 and we have:

$$N(\sigma_r^2 + \bar{r}^2) = N - z_0 s_1 s_2 2^{n-3} \sqrt{\pi} \cdot \frac{\Gamma\left(\frac{n}{2}\right)\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}$$

$$\times\left(1 + \frac{n-1}{2}\,\frac{n-1}{2}\,\frac{\rho^2}{1\,!}\,\frac{1}{\frac{n+1}{2}} + \frac{n-1}{2}\,\frac{n+1}{2}\,\frac{n-1}{2}\,\frac{n+1}{2}\,\frac{\rho^4}{2\,!}\,\frac{1}{\frac{n+1}{2}\,\frac{n+3}{2}} + \text{etc.}\right)$$

$$= N - z_0 s_1 s_2 \sqrt{\pi}\, 2^{n-3}\, \frac{\Gamma\left(\frac{n}{2}\right)\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} F\left(\frac{n-1}{2}, \frac{n-1}{2}, \frac{n+1}{2}, \rho^2\right),$$

or by Euler's transformation:

$$\sigma_r^2 + \bar{r}^2 = 1 - \frac{z_0 s_1 s_2 \sqrt{\pi}\, 2^{n-3}}{N(1-\rho^2)^{\frac{n-3}{2}}}\,\frac{\Gamma\left(\frac{n}{2}\right)\Gamma^2\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} F(1, 1, \tfrac{1}{2}(n+1), \rho^2).$$

Substituting for N from (xx), we find:

$$\sigma_r^2 = 1 - \bar{r}^2 - (1-\rho^2)\,\frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)\Gamma\left(\frac{n+1}{2}\right)} F(1, 1, \tfrac{1}{2}(n+1), \rho^2)$$

$$= 1 - \bar{r}^2 - (1-\rho^2)\frac{n-2}{n-1} F(1, 1, \tfrac{1}{2}(n+1), \rho^2) \quad \text{..................(xxvi).}$$

With a different notation this accords with the result in *Biometrika*, Vol. XI. p. 335, Equation (xx) *bis*.

We can now write down the value of $\bar{p}_{r\sigma_1} = p'_{r\sigma_1} - \bar{r}\bar{\sigma}_1$ from Equations (xxiii)—(xxv) and we find finally, if

$$\lambda_n = \Gamma\left(\frac{n}{2}\right) \Big/ \Gamma\left(\frac{n-1}{2}\right),$$

that

$$r_{r\sigma_1} = \rho\sqrt{\frac{2\lambda_n^2}{n-1-2\lambda_n^2}}\;\frac{\left(1 - \frac{\lambda_n}{\lambda_n+1} F\left(\frac{1}{2}, \frac{1}{2}, \frac{n+1}{2}, \rho^2\right)\right)}{\sigma_r} \quad \text{...(xxvi } \textit{bis}\text{).}$$

Noting (xxiii), this reduces to:

$$r_{r\sigma_1} = \sqrt{\frac{2\lambda_n^2}{n-1-2\lambda_n^2}}\;\frac{(\rho - \bar{r})}{\sigma_r} \quad \text{.....................(xxvii),}$$

and σ_r must be found from (xxv).

As the values of $\bar{r}$ and σ_r have been already found* and the radical has been computed for our earlier work in this paper, there is not much difficulty in obtaining $r_{r\sigma_1}$. The following table gives the numerical values, and shows the degree of approximation to the usual value $\rho/\sqrt{2}$.

---

* *Biometrika*, Vol. XI. pp. 372, 379—401.

TABLE II. *Values of the Correlation of the Standard Deviation of One Variate with the Correlation of Both Variates in Samples of Different Sizes taken from a Normal Population.*

Values of ρ.

Size of Sample.	*·0*	*·1*	*·2*	*·3*	*·4*	*·5*	*·6*	*·7*	*·8*	*·9*	*1·0*
2	·0000	·0481	·0958	·1431	·1893	·2340	·2762	·3141	·3438	·3532	·0000
3	·0000	·0581	·1156	·1726	·2284	·2822	·3328	·3779	·4122	·4188	·0000
4	·0000	·0620	·1237	·1847	·2445^{+}	·3026	·3575^{+}	·4073	·4469	·4608	·0000
5	·0000	·0641	·1279	·1911	·2533	·3138	·3717	·4250^{-}	·4696	·4931	·0000
6	·0000	·0654	·1305^{+}	·1952	·2586	·3210	·3809	·4369	·4857	·5179	·4134
7	·0000	·0663	·1323	·1979	·2626	·3260	·3874	·4455^{-}	·4976	·5368	·5069
8	·0000	·0669	·1336	·1998	·2653	·3297	·3922	·4519	·5067	·5512	·5550^{-}
9	·0000	·0674	·1345^{+}	·2013	·2674	·3325^{-}	·3959	·4569	·5137	·5624	·5845^{+}
10	·0000	·0677	·1353	·2025^{-}	·2691	·3347	·3989	·4609	·5194	·5712	·6045^{+}
11	·0000	·0680	·1359	·2034	·2704	·3365^{-}	·4013	·4642	·5240	·5783	·6189
12	·0000	·0683	·1364	·2042	·2715^{-}	·3380	·4033	·4669	·5278	·5841	·6298
13	·0000	·0685^{-}	·1368	·2048	·2724	·3392	·4050^{-}	·4692	·5310	·5888	·6383
14	·0000	·0686	·1372	·2054	·2732	·3403	·4064	·4712	·5338	·5928	·6452
15	·0000	·0688	·1375^{-}	·2059	·2739	·3412	·4077	·4729	·5361	·5962	·6508
16	·0000	·0689	·1377	·2063	·2744	·3420	·4088	·4743	·5381	·5992	·6554
17	·0000	·0690	·1380	·2066	·2750^{-}	·3427	·4097	·4756	·5399	·6017	·6594
18	·0000	·0691	·1382	·2069	·2754	·3434	·4106	·4767	·5414	·6040	·6628
19	·0000	·0692	·1383	·2072	·2758	·3439	·4113	·4777	·5428	·6058	·6657
20	·0000	·0693	·1385^{+}	·2075^{-}	·2762	·3444	·4120	·4786	·5441	·6076	·6683
21	·0000	·0694	·1386	·2077	·2765^{-}	·3448	·4127	·4795^{-}	·5452	·6092	·6706
22	·0000	·0694	·1388	·2079	·2768	·3452	·4131	·4802	·5462	·6105^{+}	·6726
23	·0000	·0695^{-}	·1389	·2081	·2771	·3456	·4136	·4809	·5470	·6118	·6744
24	·0000	·0695^{+}	·1390	·2083	·2773	·3460	·4141	·4815^{-}	·5479	·6130	·6761
25	·0000	·0696	·1391	·2085^{-}	·2775^{+}	·3463	·4145^{-}	·4820	·5487	·6140	·6775^{+}
50	·0000	·0701	·1403	·2104	·2802	·3500^{+}	·4195^{+}	·4886	·5576	·6258	·6936
100	·0000	·0704	·1409	·2112	·2815^{+}	·3518	·4219	·4919	·5617	·6314	·7007
400	·0000	·0706	·1413	·2119	·2825^{+}	·3531	·4237	·4942	·5647	·6350^{+}	·7055
$\rho/\sqrt{2}$	·0000	·0707	·1414	·2121	·2828	·3536	·4243	·4950	·5657	·6364	·7071

It will be clear to the reader that this table possesses more singularities than that for $r_{\sigma_1\sigma_2}$. The maximum value for $r_{\sigma_1\sigma_2}$ for a sample of definite size n may be reached only at $\rho = 1·0$, or it may in the case of samples of 8 or less be reached earlier. Further it may be doubted whether for practical purposes we can use the ordinary value, i.e. that for "large" samples $(\rho/\sqrt{2})$ at $n = 25$; certainly only roughly for $\rho = ·7$ and over. We might possibly for *practical* purposes fix the limit of small samples at $n = 50$. If we want, however, the probable error of such a simple constant as the regression coefficient, $R = r\sigma_2/\sigma_1$, for a small sample it is needful actually to know not only $\{\delta r\,\delta\sigma_1\}$, $\{\delta r\,\delta\sigma_2\}$, $\{\delta\sigma_1\,\delta\sigma_2\}$, but we clearly require higher products such as $\{\delta\sigma_1^2\}$, $\{\delta r\,\delta\sigma_1\,\delta\sigma_2\}$, $\{\delta\sigma_1\,\delta\sigma_2^2\}$, etc. and also the fourth order products to obtain σ_R^2 to what is after all only then a second order approximation. If we go to the roughest form of approximation only, we have*:

$$\sigma_R^2 = \breve{R}^2\left(\frac{\sigma_r^2}{\breve{\rho}^2} + \frac{\sigma^2_{\sigma_1}}{\breve{\sigma}_1^2} + \frac{\sigma^2_{\sigma_2}}{\breve{\sigma}_2^2} - \frac{2\{\delta\sigma_1\delta\sigma_2\}}{\breve{\sigma}_1\breve{\sigma}_2} - \frac{2\{\delta\sigma_1\delta r\}}{\breve{\sigma}_1\breve{\rho}} + \frac{2\{\delta\sigma_2\delta r\}}{\breve{\sigma}_2\breve{\rho}}\right) \text{...(xxviii).}$$

---

* A short quantity ˘ over the letter indicates that the value should be that of the population sampled.

Now such an approximation will show us how far the use of large sample values for σ_r, σ_{σ_1}, σ_{σ_2}, $\{\delta\sigma_1\,\delta\sigma_2\}$, $\{\delta\sigma_1\,\delta r\}$, $\{\delta\sigma_2\delta r\}$ modifies the customary value of σ_R.

To illustrate this let us suppose $\rho = \cdot 6$ and the sample to be of size 10. Then the value of σ_R for a large sample would be $\breve{R}\sqrt{1-\rho^2}/(\rho\sqrt{N})$[*]. This equals for our case $\cdot 4216\,\breve{R}$.

We will now insert in the above formula the values of the several terms:

$$\frac{\sigma_r}{\breve{\rho}} = \frac{\cdot 2355}{\cdot 6} = \cdot 3925\,\dagger,$$

$$\frac{\sigma_{\sigma_1}}{\sigma_1} = \cdot 2203\ddagger, \quad \frac{\sigma_{\sigma_2}}{\sigma_2} = \cdot 2203\ddagger,$$

$$\begin{aligned}
\{\delta\sigma_1\delta\sigma_2\} &= \sigma_{\sigma_1}\sigma_{\sigma_2}r_{\sigma_1\sigma_2} = \breve{\sigma}_1\breve{\sigma}_2(\cdot 2203)^2\cdot 3537 = \breve{\sigma}_1\breve{\sigma}_2\cdot 0171{,}6580,\\
\{\delta\sigma_1\delta r\ \} &= \sigma_{\sigma_1}\sigma_r\ r_{r\sigma_1} = \breve{\rho}\,\breve{\sigma}_1(\cdot 3925)(\cdot 2203)\cdot 3989 = \breve{\rho}\breve{\sigma}_1\cdot 0344{,}9199,\\
\{\delta\sigma_2\delta r\ \} &= \sigma_{\sigma_2}\sigma_r\ r_{r\sigma_2} = \breve{\rho}\,\breve{\sigma}_2(\cdot 3925)(\cdot 2203)\cdot 3989 = \breve{\rho}\breve{\sigma}_2\cdot 0344{,}9199,
\end{aligned}$$

these last three values being deduced from the present paper. Accordingly the terms in $\{\delta\sigma_1\delta r\}$ and $\{\delta\sigma_2\delta r\}$ cancel and we have

$$\begin{aligned}
\sigma_R{}^2 &= \breve{R}^2(\cdot 1540{,}5625 + 2\times\cdot 0485{,}3209 - 2\times\cdot 0171{,}6580)\\
&= \breve{R}^2\times\cdot 2167{,}8883,
\end{aligned}$$

or $$\sigma_R = \cdot 4205\breve{R}.$$

This is so close to $\cdot 4216\breve{R}$ found for σ_R from the usual formula for large samples, that it is clear that grave error would not be introduced by using the ordinary values of $\sigma_r/\breve{\rho}$ and σ_σ/σ, together with that for $\{\delta\sigma_1\delta\sigma_2\}$ in the case of $n = 10$. It is a different question, however, whether it is legitimate to neglect the contributions of the third and fourth order products to $\sigma_r{}^2$, when $1/N$ is of the order 1/10. The answer to this question cannot however be provided until these higher products have been studied and we know their numerical values for small samples. We shall return to this point in our third section.

For the special case of $r_{r\sigma_1}$ when $\rho = 1$ we require to ascertain a limit. For in this case $\bar{r}$ will equal $\rho = 1$, and σ_r will be zero so that the ratio $(\rho - \bar{r})/\sigma_r$ takes the indeterminant form 0/0. In order to evaluate the limit we require to expand $\bar{r}$ and σ_r in powers of ϵ where $\rho^2 = 1 - \epsilon$ and ϵ is made very small. First let us consider $\bar{r}$ in Equation (xxiii), we have

$$\bar{r} = \sqrt{1-\epsilon}\,\psi_n F(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n+1), 1-\epsilon),$$

where $$\psi_n = \Gamma^2\left(\frac{n}{2}\right)\Big/\left\{\Gamma\left(\frac{n-1}{2}\right)\cdot\Gamma\left(\frac{n+1}{2}\right)\right\}.$$

Hence $$\bar{r} = (1 - \tfrac{1}{2}\epsilon - \tfrac{1}{8}\epsilon^2 + \dots)\,\psi_n\left\{F_1 - \epsilon\left(\frac{dF}{dx}\right)_1 + \frac{\epsilon^2}{2}\left(\frac{d^2F}{dx^2}\right)_1 - \dots\right\},$$

---

[*] *Phil. Trans.* Vol. 191 A, p. 245 or *Biometrika*, Vol. IX. p. 9.

[†] *Biometrika*, Vol. XI. p. 386.

[‡] *Ibid.* Vol. X. p. 529.

where F_1 stands for $F(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}(n+1), 1)$ and $\left(\frac{d^sF}{dx^s}\right)_1$ for the sth differential coefficient of $F(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}(n+1), x)$, x being put unity after differentiation.

But it is well-known that

$$F(\alpha, \beta, \gamma, 1) = \frac{\Gamma(\gamma)\,\Gamma(\gamma-\alpha-\beta)}{\Gamma(\gamma-\alpha)\,\Gamma(\gamma-\beta)},$$

or we have

$$F_1(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n+1), 1) = \frac{\Gamma(\frac{1}{2}(n-1))\,\Gamma(\frac{1}{2}(n+1))}{\Gamma^2\left(\frac{n}{2}\right)} = \frac{1}{\psi_n}.$$

Thus $$\bar{r} = (1 - \tfrac{1}{2}\epsilon - \tfrac{1}{8}\epsilon^2 + \dots)\left\{1 - \frac{1}{F_1}\left(\frac{dF}{dx}\right)_1 \epsilon + \frac{1}{2F_1}\left(\frac{d^2F}{dx^2}\right)_1 \epsilon^2 - \dots\right\}.$$

Further it is easy to show from the differential equation of the hypergeometrical series that:

$$\frac{1}{F_1}\left(\frac{dF}{dx}\right)_1 = \frac{\alpha\beta}{\gamma-\alpha-\beta-1}, = \frac{1}{2(n-3)} \text{ in our case,}$$

$$\frac{1}{F_1}\left(\frac{d^2F}{dx^2}\right)_1 = \frac{\alpha\beta(\alpha\beta+\alpha+\beta+1)}{(\gamma-\alpha-\beta-1)(\gamma-\alpha-\beta-2)}, = \frac{9}{4}\frac{1}{(n-3)(n-5)} \text{ in our case.}$$

Hence we find

$$\rho - \bar{r} = \frac{1}{2(x-3)}\epsilon - \frac{2n-1}{8(n-3)(n-5)}\epsilon^2 + \dots,$$

and $$\bar{r} = 1 - \frac{1}{2}\frac{n-2}{n-3}\epsilon - \frac{(n-2)(n-8)}{8(n-3)(n-5)}\epsilon^2 + \dots.$$

Accordingly

$$1 - \bar{r}^2 = \frac{n-2}{n-3}\epsilon - \frac{(2n-7)(n-2)}{2(n-3)^2(n-5)}\epsilon^2 + \dots.$$

We now require

$$(1-\rho^2)\frac{n-2}{n-1}F(1, 1, \tfrac{1}{2}(n+1), \rho^2),$$

or $$\epsilon\frac{n-2}{n-1}\left\{F_2 - \epsilon\left(\frac{dF_2}{dx}\right)_1 + \tfrac{1}{2}\epsilon^2\left(\frac{d^2F_2}{dx^2}\right)_1 - \dots\right\},$$

where $F_2 = F(1, 1, \frac{1}{2}(n+1), \rho^2)$ and $\left(\frac{d^sF_2}{dx^s}\right)_1$ = the sth differential coefficient of $F = F_2(1, 1, \frac{1}{2}(n+1), x)$ with x put unity after differentiation.

But $$F_2 = \Gamma\left(\frac{n+1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\Big/\Gamma^2\left(\frac{n-1}{2}\right) = \frac{n-1}{n-3},$$

and $$\frac{1}{F_2}\left(\frac{dF_2}{dx}\right)_1 = \frac{2}{n-5}.$$

Thus from (xxvi) we have

$$\sigma_r^2 = \epsilon^2\left\{\frac{2(n-2)}{(n-3)(n-5)} - \frac{(n-2)(2n-7)}{2(n-3)^2(n-5)}\right\} + \text{etc.}$$

$$= \tfrac{1}{2}\epsilon^2\frac{(n-2)(2n-5)}{(n-3)^2(n-5)}.$$

Hence from the values provided on p. 191:

$$\underset{\rho=1}{\mathrm{L}}\frac{\rho-\bar{r}}{\sigma_r}=\sqrt{\frac{(n-5)}{2(n-2)(2n-5)}}\quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(\text{xxix}),$$

or

$$r_{r\sigma_1}\text{ for }\rho=1=\sqrt{\frac{2\lambda_n^2}{n-1-2\lambda_n^2}}\sqrt{\frac{(n-5)}{2(n-2)(2n-5)}}\quad\ldots\ldots\ldots(\text{xxx}).$$

When $n=5$ this limiting value is zero. For $n=6$ and higher values, it is always real and positive. For $n=2, 3$ and 4 this method of approaching the limit fails—the values are imaginary. These three cases require separate investigation.

In the case of $n=2$, $\bar{r}=\frac{2}{\pi}\sin^{-1}\rho$ * and if $\rho=1-\lambda$, we have:

$$\sin^{-1}\rho=\frac{\pi}{2}-2\sin^{-1}\sqrt{\frac{\lambda}{2}}.$$

Therefore

$$\bar{r}=1-\frac{4}{\pi}\left\{\left(\frac{\lambda}{2}\right)^{\frac{1}{2}}+\tfrac{1}{6}\left(\frac{\lambda}{2}\right)^{\frac{3}{2}}+\ldots\right\},$$

$$\rho-\bar{r}=\lambda+\frac{4}{\pi}\left\{\left(\frac{\lambda}{2}\right)^{\frac{1}{2}}+\tfrac{1}{6}\left(\frac{\lambda}{2}\right)^{\frac{3}{2}}+\ldots\right\},$$

$$\bar{r}^2=1-\frac{8}{\pi}\left\{\left(\frac{\lambda}{2}\right)^{\frac{1}{2}}+\tfrac{1}{6}\left(\frac{\lambda}{2}\right)^{\frac{3}{2}}+\ldots\right\}+\frac{8}{\pi^2}\lambda.$$

But for $n=2$, $\sigma_r^2=1-\bar{r}^2$†, or

$$\sigma_r^2=\frac{8}{\pi}\left(\frac{\lambda}{2}\right)^{\frac{1}{2}}-\frac{8}{\pi^2}\lambda+\ldots.$$

Accordingly

$$\frac{\rho-\bar{r}}{\sigma_r}=\sqrt{2\pi}\,(\tfrac{1}{2}\lambda)^{\frac{1}{4}}+\text{etc.}\quad\ldots\ldots\ \ldots\ldots\ldots\ldots\ldots\ldots(\text{xxxi}),$$

and this is zero when $\lambda=0$. The important point is that $\bar{r}$ starts with powers of $\lambda^{\frac{1}{2}}$ and not with integer powers.

We next turn to the case of $n=3$. Here we have‡:

$$\left.\begin{aligned}\bar{r}&=\frac{1}{\rho}(E_1-(1-\rho^2)F_1)\\ \sigma_r^2&=1-\bar{r}^2+\tfrac{1}{2}\cdot\frac{1-\rho^2}{\rho}\log_e(1-\rho^2)\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots\ldots(\text{xxxii}),$$

where E_1 and F_1 are the complete elliptic integrals for $k^2=\rho^2$. The difficulty of the limit here is that as ρ approaches unity E_1 approaches unity and F_1 approaches infinity. But without proceeding to a limit we have by the tables of the complete elliptic integrals:

$$\rho=\cdot 95,\qquad \bar{r}=\cdot 894{,}94275\S,\qquad \sigma_r=\cdot 270{,}7989\S,$$

$$\rho=\cdot 999{,}8477,\qquad \bar{r}=\cdot 999{,}2484\quad,\qquad \sigma_r=\cdot 016{,}41104,$$

$$\rho=\cdot 999{,}9985,\qquad \bar{r}=\cdot 999{,}987{,}813,\qquad \sigma_r=\cdot 002{,}30196,$$

---

* *Biometrika*, Vol. XI. p. 360. † *Ibid.* p. 361. ‡ *Ibid.* p. 368.

§ Erroneous numerical values for these constants are given, *Biometrika*, Vol. XI. p. 368.

giving respectively

$$r_{r\sigma} = \cdot 38895, \quad = \cdot 06986 \text{ and } = \cdot 01013,$$

showing the rapid approach to zero of the value of $r_{r\sigma}$ as ρ approaches unity.

We next turn to the case of $n = 4$.

Here if $\rho = \sin \alpha$, then

$$\left.\begin{aligned} \bar{r} &= \frac{2}{\pi}\{\cot \alpha + \alpha(1 - \cot^2 \alpha)\} \\ \text{and} \qquad \sigma_r^2 &= 1 - \bar{r}^2 - 2\cot^2 \alpha + 2\alpha \cot^3 \alpha \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots \text{(xxxiii)}.$$

Now let $\alpha = \frac{\pi}{2} - \gamma$, where γ is a small quantity. Then

$$\rho = \cos \gamma = 1 - \frac{\gamma^2}{2} + \ldots,$$

$$\begin{aligned} \bar{r} &= \frac{2}{\pi}\left\{\tan \gamma + \left(\frac{\pi}{2} - \gamma\right)(1 - \tan^2 \gamma)\right\} \\ &= \frac{2}{\pi}\left\{\gamma + \tfrac{1}{3}\gamma^3 + \left(\frac{\pi}{2} - \gamma\right)(1 - \gamma^2 + \ldots)\right\} \\ &= \frac{2}{\pi}\left(\tfrac{1}{3}\gamma^3 + \frac{\pi}{2} - \frac{\pi}{2}\gamma^2 + \gamma^3\right) \\ &= 1 - \gamma^2 + \frac{8}{3\pi}\gamma^3 + \ldots. \end{aligned}$$

Accordingly

$$1 - \bar{r}^2 = 2\gamma^2 - \frac{16}{3\pi}\gamma^3 + \ldots,$$

$$\rho - \bar{r} = \tfrac{1}{2}\gamma^2 - \text{terms in } \gamma^3,$$

$$\begin{aligned} \sigma_r^2 &= 1 - \bar{r}^2 - 2\tan^2 \gamma + 2\left(\frac{\pi}{2} - \gamma\right)\tan^3 \gamma \\ &= \left(\pi - \frac{16}{3\pi}\right)\gamma^3 - \tfrac{7}{3}\gamma^4 + \ldots. \end{aligned}$$

Hence

$$\underset{\rho=1}{\mathrm{L}} \frac{\rho - \bar{r}}{\sigma_r} = \underset{\rho=1}{\mathrm{L}} \frac{\tfrac{1}{2}\gamma^{\frac{1}{2}}}{\sqrt{\left(\pi - \frac{16}{3\pi}\right)}}, \text{ when } \gamma = 0 \quad \ldots\ldots\ldots\ldots \text{(xxxiv)},$$

$$= 0.$$

Or we conclude that for $n = 4$, $r_{r\sigma_1} = 0$. We thus see that for all the cases in which our limit to $(\rho - \bar{r})/\sigma_r$ becomes imaginary a direct investigation shows that $r_{r\sigma_1} = 0$. We now return for a moment to the method by which our limit for $n > 5$ was obtained, and we ask: Is the expansion by Taylor's series of the hypergeometrical legitimate? The doubt which exists in the process is this: The differentials of the hypergeometrical series for the values of α, β (i.e. both $\frac{1}{2}$ or both 1) and $\gamma = \frac{1}{2}(n + 1)$ will ultimately become indefinite for some one or other value of n. Can we therefore use Taylor's theorem at all? I think we can. For if we are considering the limit when $n = s$ and accordingly a certain differential coefficient becomes infinite, say $\frac{d^u F}{dx^u}$ when $x = 1$, we have really to consider what happens to

$\epsilon^u \dfrac{d^u F}{dx^u}$ in such a case when ϵ approaches zero, and we can always make ϵ approach zero by taking it equal to $\epsilon'(n-s)$, so that the factor $n-s$ disappears from the denominator. For our purposes we need only the first two terms of Taylor's expansion. Thus, if θ be a fraction between 0 and 1, we shall have:

$$F_1\left(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}(n+1), 1-\epsilon\right)$$

$$= F_1 \times \left\{1 - \frac{1}{2(n-3)}\epsilon + \frac{9}{8(n-3)(n-5)}\left(1 - \frac{25}{2(n-7)}\epsilon^\theta\right)\epsilon^2\right\} \quad \text{...(xxxv)},$$

where the term in θ expresses the remainder after two terms. It may be said that this will become infinite when $n=7$. But as we are going to make ϵ approach zero we can take it $=(n-7)\epsilon'$, where ϵ' is also infinitely small, and we reach precisely the same limit as before. It seems therefore that for $n>5$ our method of approaching the limit will apply. At the same time a full discussion of the subject by a more competent mathematician is desirable.

III. *On the Standard Deviation of the Regression Coefficient, $r\sigma_1/\sigma_2$, and of the Standard Deviation of Arrays, $\sigma_1\sqrt{1-r^2}$, in samples drawn from a normal population.*

As before the frequency surface is

$$z = z_0' e^{-\frac{1}{2}\left(\frac{\sigma_1}{s_1}\right)^2} e^{-\frac{1}{2}\left(\frac{\sigma_2}{s_2}\right)^2} \left(\frac{\sigma_1\sigma_2}{s_1 s_2}\right)^{n-2} e^{\gamma r}(1-r^2)^{\frac{n-4}{2}} \quad \text{.........(xxxvi)};$$

we need to multiply this in succession by

$$r\sigma_1/\sigma_2, \quad (r\sigma_1/\sigma_2)^2, \quad \sigma_1\sqrt{1-r^2}, \text{ and } \sigma_1^2(1-r^2),$$

and integrate over the whole volume, i.e. $r=1$ to -1, $\sigma_1 = 0$ to ∞, $\sigma_2 = 0$ to ∞.

We can do this most expeditiously by evaluating the integral of z multiplied by $r^p(1-r^2)^{\frac{1}{2}t}\sigma_1^{q_1}\sigma_2^{q_2}$.

$$W_{p,t,q_1,q_2} = \int_{-1}^{+1}\int_0^\infty\int_0^\infty z r^p \sigma_1^{q_1}\sigma_2^{q_2}(1-r^2)^{\frac{1}{2}t}\,dr\,d\sigma_1\,d\sigma_2$$

$$= z_0' s_1^{q_1+1} s_2^{q_2+1}\int_0^\infty\int_0^\infty e^{-\frac{1}{2}(\sigma_1'^2+\sigma_2'^2)}\,\sigma_1'^{n+q_1-2}\sigma_2'^{n+q_2-2} \times \int_{-1}^{+1} e^{\gamma r} r^p (1-r^2)^{\frac{1}{2}(n+t-4)}\,dr\,d\sigma_1\,d\sigma_2$$

$$\text{...........(xxxvii)},$$

where for brevity σ_1' and σ_2' are written for σ_1/s_1 and σ_2/s_2. Let

$$I_{p,t} = \int_{-1}^{+1} e^{\gamma r} r^p (1-r^2)^{\frac{1}{2}(n+t-4)}\,dr \quad \text{.................(xxxviii)}.$$

We have to distinguish p odd from p even.

(i) p odd. Take $r=\sin\theta$ and expand the exponential; only the odd powers of γ remain, and

$$I_{p,t} = 2\int_0^{\frac{\pi}{2}}\left(\gamma\sin^{p+1}\theta + \frac{\gamma^3}{3!}\sin^{p+3}\theta + \dots\right)\cos^{n+t-3}\theta\,\delta\theta.$$

Put $x = \sin^2\theta$, and we have:

$$I_{p,t} = \int_0^1 \left(\gamma x^{\frac{1}{2}p} + \frac{\gamma^3}{3!} x^{\frac{1}{2}(p+2)} + \ldots\right)(1-x)^{\frac{1}{2}(n+t-4)}\,dx$$

$$= \frac{\Gamma(\frac{1}{2}(p+2))\,\Gamma(\frac{1}{2}(n+t-2))}{\Gamma(\frac{1}{2}(p+n+t))}\left(\frac{\gamma}{1!} + \frac{\gamma^3}{3!}\frac{p+2}{p+n+t} + \frac{\gamma^5}{5!}\frac{(p+2)(p+4)}{(p+n+t)(p+n+t+2)} + \ldots\right) \qquad \text{(xxxix)}.$$

(ii) p even. We easily find:

$$I_{p,t} = 2\int_0^{\frac{\pi}{2}} \left(\sin^p\theta + \frac{\gamma^2}{2!}\sin^{p+2}\theta + \ldots\right)\cos^{n+t-3}\theta\, d\theta$$

$$= \int_0^1 \left(x^{\frac{1}{2}(p-1)}\theta + \frac{\gamma^2}{2!}x^{\frac{1}{2}(p+1)}\theta + \ldots\right)(1-x)^{\frac{n+t-4}{2}}\,dx$$

$$= \frac{\Gamma(\frac{1}{2}(p+1))\,\Gamma(\frac{1}{2}(n+t-2))}{\Gamma(\frac{1}{2}(p+n+t-1))}\left(1 + \frac{\gamma^2}{2!}\frac{p+1}{p+n+t-1} + \frac{\gamma^4}{4!}\frac{(p+1)(p+3)}{(p+n+t-1)(p+n+t+1)} + \ldots\right) \qquad \text{(xl)}.$$

Now $\gamma = \rho\sigma_1\sigma_2/(s_1 s_2) = \rho\sigma_1'\sigma_2'$. Hence when we come to integrate these two expressions for σ_1' and σ_2' we require the integral, in the case of γ^u, of

$$U = \rho^u \int_0^\infty \int_0^\infty e^{-\frac{1}{2}\sigma_1'^2} e^{-\frac{1}{2}\sigma_2'^2}\, \sigma_1'^{\,n+u+q_1-2}\, \sigma_2'^{\,n+u+q_2-2}\, d\sigma_1' d\sigma_2'.$$

Changing the integration to y_1 and y_2 by the relations $\sigma_1'^2 = 2y_1$, $\sigma_2'^2 = 2y_2$, we have

$$U = \rho^u 2^{n+u+\frac{1}{2}(q_1+q_2)-3} \int_0^\infty e^{-y_1} y_1^{\frac{1}{2}(n+u+q_1-3)}\, dy_1 \times \int_0^\infty e^{-y_2} y_2^{\frac{1}{2}(n+u+q_2-3)}\, dy_2,$$

or

$$U = \rho^u 2^{n+u+\frac{1}{2}(q_1+q_2)-3}\, \Gamma(\tfrac{1}{2}(n+u+q_1-1))\,\Gamma(\tfrac{1}{2}(n+u+q_2-1)) \qquad \text{(xli)}.$$

We have now to apply (xxxix) to (xli) to our special cases.

(a) *Mean value of* $r\sigma_1/\sigma_2 = \{r\sigma_1/\sigma_2\}$, *say.*

Here $p = 1$, $q_1 = 1$, $q_2 = -1$, $t = 0$. We have:

$$N\{r\sigma_1/\sigma_2\} = z_0' s_1^2 \frac{\Gamma(\frac{3}{2})\,\Gamma(\frac{1}{2}(n-2))}{\Gamma(\frac{1}{2}(n+1))} 2^{n-3}\, \Gamma(\tfrac{1}{2}(n+1))\,\Gamma(\tfrac{1}{2}(n-1))$$

$$\times \left(2\rho + \frac{\rho^3}{3!}\frac{3}{n+1} 2^3 \frac{n+1}{2}\frac{n-1}{2} + \frac{\rho^5}{5!} 2^5 \frac{3.5}{(n+1)(n+3)} \frac{n+1}{2}\frac{n+3}{2}\frac{n-1}{2}\frac{n+1}{2} + \ldots\right)$$

$$= z_0' s_1^2\, 2^{n-3} \sqrt{\pi}\, \Gamma(\tfrac{1}{2}(n-2))\,\Gamma(\tfrac{1}{2}(n-1)) \frac{\rho}{(1-\rho^2)^{\frac{1}{2}(n-1)}} \qquad \text{(xlii)}.$$

Hence, using (xix),

$$\{r\sigma_1/\sigma_2\} = \rho s_1/s_2 = \rho\Sigma_1/\Sigma_2 \qquad \text{(xliii)},$$

which is the expression for the mean slope of the regression lines of all samples. But as r is *not* symmetrically distributed about the value ρ and the mean value of r, i.e. $\bar{r}$, is not ρ (see (xxiii)), we could not *a priori* assume that the mean value of $(r\sigma_1/\sigma_2)$ would be $\rho\Sigma_1/\Sigma_2$, the slope of the regression line in the sampled population.

(*b*) *Mean value of* $(r\sigma_1/\sigma_2)^2 = \{(r\sigma_1/\sigma_2)^2\}$, *say*.

Here we must put $p = 2$, $q_1 = 2$, $q_2 = -2$, $t = 0$. We have:

$$U = \rho^u 2^{n+u-3}\,\Gamma\left(\tfrac{1}{2}(n+u+1)\right)\Gamma\left(\tfrac{1}{2}(n+u-3)\right),$$

and

$$N\{(r\sigma_1/\sigma_2)^2\} = z_0'\,\frac{s_1^3}{s_2}\,\frac{\Gamma(\frac{3}{2})\,\Gamma(\frac{1}{2}(n-2))}{\Gamma(\frac{1}{2}(n+1))}\,2^{n-3}\,\Gamma\left(\tfrac{1}{2}(n+1)\right)\Gamma\left(\tfrac{1}{2}(n-3)\right)$$

$$\times\left[1 + \frac{3\rho^2}{1!}\,\frac{n-3}{2} + \frac{5\rho^4}{2!}\,\frac{n-3}{2}\,\frac{n-1}{2}\,\frac{7\rho^6}{3!} + \frac{n-3}{2}\,\frac{n-1}{2}\,\frac{n+1}{2} + \dots\right].$$

But the series on the right in square brackets is

$$\frac{d}{d\rho}\left(\rho\,(1-\rho^2)^{-\frac{1}{2}(n-3)}\right) = \frac{1+(n-4)\,\rho^2}{(1-\rho^2)^{\frac{1}{2}(n-1)}}.$$

Thus, using (xix) again:

$$\{(r\sigma_1/\sigma_2)^2\} = \tfrac{1}{2}\left(\frac{s_1}{s_2}\right)^2\frac{\Gamma(\frac{1}{2}(n-3))}{\Gamma(\frac{1}{2}(n-1))}\,(1+(n-4)\,\rho^2) \quad\text{............(xliv)},$$

and, using R to denote $r\sigma_1/\sigma_2$:

$$\begin{aligned}\sigma_R^2 &= \{(r\sigma_1/\sigma_2)^2\} - \{r\sigma_1/\sigma_2\}^2\\ &= \frac{\Sigma_1^2}{\Sigma_2^2}\left(\frac{\Gamma(\frac{1}{2}(n-3))}{2\Gamma(\frac{1}{2}(n-1))}(1+(n-4)\,\rho^2) - \rho^2\right)\\ &= \frac{\Sigma_1^2}{\Sigma_2^2}\left(\frac{1+(n-4)\,\rho^2}{n-3} - \rho^2\right) = \frac{\Sigma_1^2}{\Sigma_2^2}\,\frac{1-\rho^2}{n-3}.\end{aligned}$$

Accordingly $$\sigma_R = \frac{1}{\sqrt{n-3}}\,\frac{\Sigma_1}{\Sigma_2}\,\sqrt{1-\rho^2} \quad\text{..................... ......(xlv)}.$$

Now the value usually taken for large samples* is:

$$\sigma_R = \frac{1}{\sqrt{n}}\,\frac{\Sigma_1}{\Sigma_2}\,\sqrt{1-\rho^2} \quad\text{..............................(xlvi)}.$$

Thus the error in applying this to small samples lies solely in using $\frac{1}{\sqrt{n}}$ for $\frac{1}{\sqrt{n-3}}$. For the case of $\rho = \cdot 6$ and $n = 10$ (see our p. 190), the true value of $\sigma_R = \cdot 5040\breve{R}$, while (xlvi) gave $\cdot 4216\breve{R}$ and the method of our p. 190, $\cdot 4205\breve{R}$. The error therefore in applying (xlvi) to compute σ_R does not lie in using the wrong values $\{\delta\sigma_1\delta\sigma_2\}$, $\{\delta r\,\delta\sigma_1\}$, $\{\delta r\,\delta\sigma_2\}$ for small samples, but in neglecting terms in higher products in (xxviii). The formula (xlv) is as simple as that for large samples, and I have not considered it needful to provide a table. The value under the radical is somewhat unexpected and suggests that special treatment is required for $n = 3$. In this case we have really to find the integral of $\frac{\sigma_2}{s_2}\,e^{\rho r\frac{\sigma_1\sigma_2}{s_1 s_2}}\,e^{-\frac{1}{2}\left(\frac{\sigma_2}{s_2}\right)^2}$ and the limits being 0 and ∞, we approach at the lower limit the infinite area of a rectangular hyperbola. The fact is that for $n = 3$, the frequency of cases of $\sigma_2 = 0$ is finite, and the second moment of the regression coefficient becomes infinite. The case is still more exaggerated for $n = 2$. Hence we conclude that it is not possible to determine the mean slope of the regression line for samples of less size than four.

---

* *Phil. Trans.* Vol. 191 A, p. 245 or *Biometrika*, Vol. IX. p. 9.

(c) *Mean value of* $\sigma_1 \sqrt{1 - r^2} = \{\sigma_a\}$, *say.*

Here $p = 0$, $q_1 = 1$, $q_2 = 0$, $t = 1$,

$$U = \rho^u \, 2^{n+u-s+\frac{1}{2}} \, \Gamma(\tfrac{1}{2}(n + u)) \, \Gamma(\tfrac{1}{2}(n + u - 1)),$$

and we find:

$$N\sigma_a = z_0' s_1^2 s_2 \frac{\Gamma(\frac{1}{2}) \, \Gamma(\frac{1}{2}(n-1))}{\Gamma(\frac{1}{2}n)} \sqrt{2} \, 2^{n-3} \, \Gamma(\tfrac{1}{2}n) \, \Gamma(\tfrac{1}{2}(n-1))$$

$$\times \left(1 + \frac{\rho^2}{1!} \frac{n-1}{2} + \frac{\rho^4}{2!} \frac{n-1}{2} \frac{n+1}{2} + \frac{\rho^6}{3!} \frac{n-1}{2} \frac{n+1}{2} \frac{n+3}{2} + \dots\right),$$

and this series is again $(1 - \rho^2)^{-\frac{1}{2}(n-1)}$.

Hence using (xix) we find:

$$\begin{aligned} \{\sigma_a\} &= \sqrt{2} \, s_1 \, \Gamma(\tfrac{1}{2}(n-1)/\Gamma(\tfrac{1}{2}(n-2)) \\ &= \Sigma_1 \sqrt{1 - \rho^2} \frac{\Gamma(\frac{1}{2}(n-1))}{(n/2)^{\frac{1}{2}} \, \Gamma(\frac{1}{2}(n-2))} \\ &= \Sigma_1 \sqrt{1 - \rho^2} \, \lambda_{n-1} / \sqrt{\tfrac{1}{2}n} \quad \dots\dots\dots\dots (\text{xlvii}), \end{aligned}$$

in our previous notation. Thus the mean value of the standard deviation of arrays is *not* the value in the sampled population, until n is so large that we can take $\lambda_{n-1}/\sqrt{\frac{1}{2}n}$ as substantially unity. In this matter the mean of the array standard deviation differs from the mean of the slope of the regression line. I give a table by which the approach of $\{\sigma_a\}$ to σ_A, the standard deviation of arrays in the sampled population, can be judged.

TABLE III.

Ratio of Mean S.D. of Arrays in Samples to Array S.D. in Sampled Population.

Size of Sample	Ratio $\{\sigma_a\}/\sigma_A$	Size of Sample	Ratio $\{\sigma_a\}/\sigma_A$	Size of Sample	Ratio $\{\sigma_a\}/\sigma_A$	Size of Sample	Ratio $\{\sigma_a\}/\sigma_A$
2	·0000	9	·8511	16	·9189	23	·9442
3	·4607	10	·8670	17	·9238	24	·9466
4	·6267	11	·8798	18	·9282	25	·9488
5	·7136	12	·8904	19	·9321	50	·9747
6	·7675^{-}	13	·8992	20	·9356	100	·9874
7	·8042	14	·9068	21	·9388	400	·9969
8	·8308	15	·9132	22	·9416		

Clearly in this case we can hardly speak of "small samples" ceasing—for practical purposes even—before $n = 50$ to 100.

(*d*) *Mean value of $\sigma_1^2 (1 - r^2)$ and standard deviation of the standard deviation of arrays in samples.*

Here $p = 0$, $q_1 = 2$, $q_2 = 0$, $t = 2$; and

$$U = \rho^u \, 2^{n+u-2} \, \Gamma(\tfrac{1}{2}(n+u+1)) \, \Gamma(\tfrac{1}{2}(n+u-1)).$$

Hence $$N\{\sigma_1^2(1-r^2)\} = z_0' \, s_1^3 s_2 \sqrt{\pi} \, 2^{n-2} \, \Gamma(\tfrac{1}{2}n) \, \Gamma(\tfrac{1}{2}(n-1))$$
$$\times \left(1 + \frac{\rho^2}{1!}\frac{n-1}{2} + \frac{\rho^4}{2!}\frac{n-1}{2}\frac{n+1}{2} + \frac{\rho^6}{3!}\frac{n-1}{2}\frac{n+1}{2}\frac{n+3}{2} + \dots\right),$$

and the series $= (1-\rho^2)^{-\frac{1}{2}(n-1)}$.

We have using (xix):

$$\{\sigma_1^2(1-r^2)\} = s_1^2 \, \frac{2\,\Gamma(\tfrac{1}{2}n)\,\Gamma(\tfrac{1}{2}(n-1))}{\Gamma(\tfrac{1}{2}(n-2))\,\Gamma(\tfrac{1}{2}(n-1))}$$
$$= s_1^2(n-2),$$

or $$\{\sigma_1^2(1-r^2)\} = \Sigma_1^2(1-\rho^2)\,\frac{n-2}{n} \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots (\text{xlviii}).$$

Thus the mean variance of the arrays in samples is not the variance in arrays of the sampled population, but the latter multiplied by the factor $(n-2)/n$.

To find the standard deviation of the standard deviation of arrays, we have:

$$\sigma^2_{\sigma_a} = \{\sigma_1^2(1-r^2)\} - \{\sigma_1\sqrt{1-r^2}\}^2$$
$$= \Sigma_1^2(1-\rho^2) \, . \left(\frac{n-2}{n} - \frac{2\lambda^2_{n-1}}{n}\right),$$

or $$\sigma_{\sigma_a} = \frac{\Sigma_1\sqrt{1-\rho^2}}{\sqrt{2n}}\,(2n - 4 - 4\lambda^2_{n-1})^{\frac{1}{2}} \quad \dots\dots\dots\dots\dots\dots (\text{xlix}).$$

Clearly $\Sigma_1\sqrt{1-\rho^2}/\sqrt{2n}$ is the value we should have given to σ_{σ_a} on the theory of large samples. Calling this value $(\sigma_{\sigma_a})_L$ I have computed the values of the ratio $\sigma_{\sigma_a}/(\sigma_{\sigma_a})_L$ and they are given in Table IV.

TABLE IV.

Values of the Ratio $\sigma_{\sigma_a}/(\sigma_{\sigma_a})_L$ for Various Sized Samples.

Size of Sample	Ratio	Size of Sample	Ratio	Size of Sample	Ratio	Size of Sample	Ratio
2	·0000	*9*	·9808	*16*	·9907	*23*	·9939
3	·8525+	*10*	·9834	*17*	·9914	*24*	·9942
4	·9265+	*11*	·9853	*18*	·9919	*25*	·9944
5	·9524	*12*	·9868	*19*	·9924	*50*	·9974
6	·9651	*13*	·9881	*20*	·9928	*100*	·9987
7	·9725+	*14*	·9891	*21*	·9932	*400*	·9997
8	·9774	*15*	·9900	*22*	·9936		

It is clear that for most practical purposes we may suppose that for a sample of 25 our ratio may be taken as unity.

It must be confessed of course that the results worked out in this paper only cover a limited series of sampled populations, i.e. those which follow the normal law. Their value is therefore rather for the light they throw on the general nature of small samples than on their direct application to any given series of small samples. We are a long way yet from a determination of the variation limits in the case of bi-variate small samples from *any* indefinitely large population whatever its law of distribution. Even in the present cases it is seen that the frequency distributions of the constants in samples have very complicated surfaces* of which it is not easy to get a clear geometrical conception.

For much help in the calculation of the tables I have to thank my colleagues Miss Ethel M. Elderton and Miss M. Moul; for computing the λ_n functions I am indebted to Miss B. N. Stoessiger, a statistical student in the Laboratory. Without their assistance the publication of this paper would have been indefinitely delayed, and I am most grateful for the aid they have given in the matter.

---

* There is some hope of expressing by a finite function the distribution of regression coefficients. But in this case the higher moments of the distribution will ultimately become infinite for small samples.

[While preparing this memoir from my notes, I received a paper from Professor Romanovsky. Writing without knowledge of the papers in *Biometrika* by "Student," Fisher, myself and others (Vol. VI. pp. 1—25, Vol. X. pp. 507—521, Vol. X. pp. 522—529, and Vol. XI. pp. 328—417) and naturally without knowledge of my present paper, Professor Romanovsky had reached, dealing only with the algebraical side, many of the published results and certain additional ones. While willing to publish the latter, the present cost of printing prohibited the reproduction of much work already published or about to be published in this *Journal*. Recognizing the difficulty of an Editor in such matters, I sent him a proof of this paper and asked him to cable if he were willing that I should add under the title his name to my own. I have at last received a letter from Professor Romanovsky. He is satisfied with the statement that many results contained in the present and earlier papers have also been obtained by him quite independently and by a different method. I trust for the sake of his additional results that his paper may shortly be published elsewhere. This note explains the delay in the issue of the present parts of *Biometrika*. K. P. *June* 25, 1925.]

Values of the Differences of the Powers of Zero.

Table of $q\,(p, s) = \dfrac{\Delta^p 0^{p+s}}{\Gamma\,(p+s+1)}$ *from* $p = 1$ *to* 10 *and* $s = 0$ *to* 10.

Values of *p*.

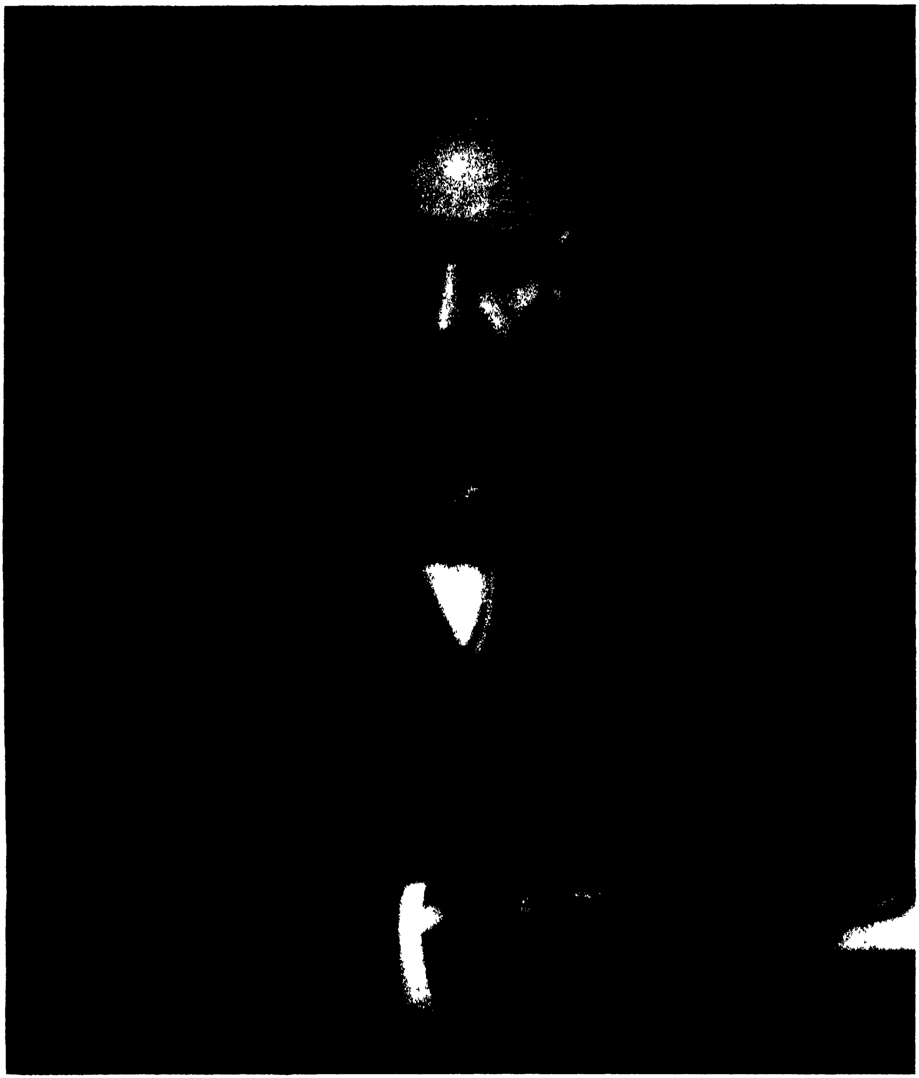
Values of *s*.

s	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*s*
0	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	1·000,000,000,000	
1	·500,000,000,000	1·000,000.000,000	1·500,000,000,000	2·000,000,000,000	2·500,000,000,000	3·000,000,000,000	3·500,000,000,000	4·000,000,000,000	4·500,000,000,000	5·000,000,000,000	
2	·1666,66,666,667	·583,333,333,333	1·250,000,000,000	2·166,666,666,667	3·333,333,333,333	4·750,000,000,000	6·416,666,666,667	8·333,333,333,333	10·500,000,000,000	12·916,666,666,667	
3	·041,666,666,667	·250,000,000,000	·750,000,000,000	1·666,666,666,667	3·125,000,000,000	5·250,000,000,000	8·166,666,666.667	12·000,000,000,000	16·875,000,000,000	22·916,666,666,667	
4	·008,333,333,333	·086,111,111,111	·358,333,333,333	1·012,500,000,000	2·298,611,111,111	4·529,166,666,667	8·079,166,666,667	13·386,111,111,111	20·950,000,000,000	31·333,333,333,333	
5	·001,388,888,889	·025,000,000,000	·143,750,000,000	·513,888,888,889	1·406,250,000,000	3·237,500,000.000	6·601,388,888,889	12·300,000,000,000	21·375,000,000,000	35·138,888,888,889	
6	·000,198,412,698	·006,299,603,175	·050,016,534,392	·225,562,169,312	·741,732,804,233	1·989,616,402,116	4·625,925,925,926	9·671,957,671,958	18·628,174,603,175-	33·604.414,682,540	
7	·000,024,801,587	·001,405,423,280	·015,426,587,302	·087,632,275,132	·345,568,783,069	1·077,777,777,778	2·851,851,851,852	6·679,365,079,365	14·235,491,071,429	28·141,121,031,746	
8	·000,002,755,732	·000,281,635,802	·004,284,060,847	·030,638,778,660	·144,695,216,049	·523,916,997,355	1·575,358,796,296	4·127,361,937,831	9·721,510,416,667	21·034,795,249,118	
9	·000,000,275,573	·000,051,256,614	·001,083,829,365	·009,760,802,469	·055,162.863.757	·231,631,944.444	·790,558,449,074	2·314,315,476,190	6·017,912,946,431	14·238,267,471,342	
10	·000,000,025,052	·000,008,546,944	·000,252,086,841	·002,860,825,517	·019,341,229,758	·094.114.940,326	·364,277,277.988	1·190,485,668,525-	3·414,504,607,081	8·826,386,039,207	*1*

This table was computed by K. P. and E. M. E. from the formula $q\,(p, s) = \dfrac{p}{p+s}\,\{q\,(p, s-1) + q\,(p-1, s)\}$, and checked by the formula

$$\Delta^p 0^{p+10} = p^{p+10} - p\,(p-1)^{p+10} + \frac{p\,(p-1)}{2\,!}\,(p-2)^{p+10} - \ldots + (-1)^r\,\frac{p!}{(p-r)\,!\,r\,!}\,(p-r)^{p+10} + \ldots .$$

It was calculated for a special investigation, but it is thought that it may be of value to other computers, and it is accordingly published here. There may be a doubt as to the twelfth decimal in the latter part of the table, but the tenth row checked properly, and isolated values tested here and there in the table also checked. As far as the computers are aware, no table has hitherto been published showing the ratio of the differences of zero to the factorial.



Rud. Martin

Rudolf Martin, 1864—1925. "Lehrbuch der Anthropologie," 1914.

"*Biometrika*" *Portrait Series*, No. II. Issued with Vol. XVII, Parts III and IV.

VOLUME XVII DECEMBER, 1925 NOS. 3 AND 4

JAMES BERNOULLI'S THEOREM.

BY KARL PEARSON.

(1) IN the course of lectures on the history of statistics given during the last three or four years, I have felt it needful to consult original papers and have been struck by the manner in which the history of science gets created and attributions accepted. Instances of this may be found in such widespread notions as that Leibnitz was the first to use differentials, that Gauss was the discoverer of the normal curve of errors, that Lagrange invented Lagrange's formula, that Bessel originated Bessel's method of interpolation, that Bravais introduced the coefficient of correlation, or that James Bernoulli demonstrated James Bernoulli's Theorem. It is with the latter point that I wish to concern myself in the present paper. In all the French and German works on probability with which I am acquainted, there is a chapter devoted to "James Bernoulli's Theorem." The statement of this theorem is usually summed up in some such phrase as this: Accuracy increases with the square root of the number of observations. And the fact that the constants of frequency distributions in the case of *large* samples have standard deviations varying inversely as the square root of the size of the sample is repeatedly spoken of as an illustration of Bernoulli's Principle, or as part of the generalised theorem of Bernoulli. The proofs provided of Bernoulli's Theorem in the text-books referred to can be traced with variations—for the majority of text-book writers copy their predecessors as well in mathematics as in anatomy—back to Laplace, and through him to De Moivre's second supplement to the *Miscellanea Analytica*, 1733.

Usually there is a more or less vague reference to the *Ars Conjectandi* of 1713. It is not, however, my purpose to trace here the tortuous process by which De Moivre's Theorem has in the course of historical evolution become associated with Bernoulli. I want rather to discuss what Bernoulli really did achieve, to investigate whether his method of approaching the subject can be profitably pushed further, and to indicate that if the text-book writers had really examined the *Ars Conjectandi* they would never have attributed to Bernoulli the law that accuracy varies with the square root of the number of observations. There is no such law and no near approach to such a law to be found in the fourth and fifth chapters of the *Pars Quarta* of Bernoulli's *Ars Conjectandi*. What Bernoulli endeavours to do is to prove that by increasing sufficiently the number of observations he can cause the probability—i.e. that the ratio of observed successful to unsuccessful occurrences will differ from the true ratio within certain small limits—to diverge from certainty by any assignable limit. Now the problem is clearly one which depends on summing any number of terms of a binomial, and therefore will be solved with complete accuracy—at any rate for the range of the arguments of the table—when the *Tables of the Incomplete* B-*Function* now under construction are published*. Approximations to the sum of terms of a binomial may be made in various ways, for example by the normal curve as was done by De Moivre for this very purpose. But James Bernoulli did not proceed in this manner, he adopted a very crude method of inequalities which I will shortly reproduce. Let us suppose that N is the true number of observations which must be made to give the required value of the probability, then clearly if m be > 1, mN will give a probability falling within the required limits. But the practical value of the solution must depend on m being nearly unity. This is far from the case with Bernoulli's solution. He gets most exaggerated values for the needful number of observations, and for this reason his solution must be said to be from the practical standpoint a failure; it would ruin either an insurance society or its clients, if it were adopted. All Bernoulli achieved was to show that by increasing the number of observations the results would undoubtedly fall within certain limits, but he failed entirely to determine what the *adequate* number of observations were for such limits. That was entirely De Moivre's discovery.

(2) *Bernoulli's Attempt at Solution.* I propose to reproduce this in its main features, without following Bernoulli verbatim.

Bernoulli considers the binomial $(r+s)^{nt}$, where $t=r+s$. The general term of this is

$$\frac{|\underline{nt}}{\underline{nt-p}\,|\underline{p}}\, r^{nt-p}\, s^{p},$$

and the greatest term is that for which $p = ns$, i.e.

$$\frac{|\underline{nt}}{|\underline{nr}\,|\underline{ns}}\, r^{nr}\, s^{ns}.$$

In the binomial the powers of r will be those in the first, the corresponding powers of s those in the second line below:

$$\begin{array}{llllllllll} nr+ns, & nr+ns-1, & \ldots & nr+n, & \ldots\; nr, & \ldots\; nr-n, & \ldots & 3,\; 2, & 1, & 0, \\ 0, & 1, & 2,\; 3 \ldots & ns-n, & \ldots\; ns, & \ldots\; ns+n, & \ldots\ldots\ldots & ns+nr-1, & ns+nr. & \end{array}$$

From the maximum term to $nr-n$ we have n terms and from the term $nr+n$ to the maximum term we have another n terms; beyond the $(nr-n)$th term we have $(r-1)\,n$ terms and before the $(nr+n)$th, $(s-1)\,n$ terms. Bernoulli proposes to consider the ratio of the sum of the n terms from $nr-1$ to $nr-n$ inclusive to the sum of the $n(r-1)$ beyond, and the sum of the n terms from $nr+n$ to $nr+1$ to the sum of the $n(s-1)$ terms which precede the $(nr+n)$th term; i.e. omitting the nrth term for the moment, he is considering the sum of the $2n$ middle terms in relation to the sum of the remaining tails. Bernoulli notes that on either side of the greatest term the terms continually decrease, and that their rate of decrease becomes more and more rapid. We shall follow Bernoulli's process a little more clearly if we write the term containing the pth power of r as u_p. Now let us suppose c a quantity such that:

---

* See *Biometrika*, Vol. xvi. p. 68.

$$c < \frac{u_{nr+1} + u_{nr+2} + \dots + u_{nr+n}}{u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n} + n(s-2) \text{ other terms}} \quad \ldots\ldots\text{(i)}.$$

Next let $$u_{nr}/u_{nr+n} = x.$$

By what precedes

$$\frac{u_{nr+1}}{u_{nr+2}} < \frac{u_{nr+n+1}}{u_{nr+n+2}}, \quad \frac{u_{nr+2}}{u_{nr+3}} < \frac{u_{nr+n+2}}{u_{nr+n+3}} \text{ and so on.}$$

Therefore:

$$\frac{u_{nr+1}}{u_{nr+n+1}} < \frac{u_{nr+2}}{u_{nr+n+2}}, \quad \frac{u_{nr+2}}{u_{nr+n+2}} < \frac{u_{nr+3}}{u_{nr+n+3}} \text{ and so on;}$$

while $\dfrac{u_{nr}}{u_{nr+n}} = x$ is less than all these.

Thus: $$xu_{nr+n+1} < u_{nr+1}, \quad xu_{nr+n+2} < u_{nr+2} \text{ and so on.}$$

Accordingly:

$$x < \frac{u_{nr+1} + u_{nr+2} + \dots + u_{nr+n}}{u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\text{(ii)}.$$

But: $$u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n} + n(s-2) \text{ other terms}$$

is $$< (s-1)(u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n}),$$

and $$\frac{x}{s-1} < \frac{u_{nr+1} + u_{nr+2} + \dots + u_{nr+n}}{(s-1)(u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n})}$$

$$< \frac{u_{nr+1} + u_{nr+2} + \dots + u_{nr+n}}{u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n} + n(s-2) \text{ other terms}} \quad \ldots\ldots\text{(iii)}.$$

Now this ratio is greater than c. If then we take $c < \dfrac{x}{s-1}$ it will follow that

$$\frac{u_{nr+1} + u_{nr+2} + \dots + u_{nr+n}}{u_{nr+n+1} + u_{nr+n+2} + \dots + u_{nr+n+n} + n(s-2) \text{ other terms}}$$

will be $> c$.

We can deal similarly with the terms beyond the maximum term, and if we choose the same value of c we shall have:

$$x > c(r-1).$$

Accordingly we have:

$$u_{nr+1} + u_{nr+2} + \dots + u_{nr+n} > c\,(u_{nr+n+1} + \dots + u_{nr+n+n} + n(s-2) \text{ other terms}),$$

$$u_{nr-1} + u_{nr-2} + \dots + u_{nr-n} > c\,(u_{nr-n-1} + \dots + u_{nr-n-n} + n(r-2) \text{ other terms}),$$

and of course:

$$u_{nr+1} + u_{nr+2} + \dots + [u_{nr}]^* + u_{nr-1} + \dots + u_{nr-n}$$

$$> c\,\{\text{sum of both remaining tails}\} \ldots\ldots\ldots\ldots\text{(iv)}.$$

---

[*] Bernoulli does not refer to the maximum term u_{nr} but it clearly only increases the inequality.

The question next to be answered is what value shall we give to x?

Now: $$x = u_{nr}/u_{nr+n} = \frac{\underline{|nt}}{\underline{|nr}\,\underline{|ns}}\, r^{nr} s^{ns} \Big/ \frac{\underline{|nt}}{\underline{|nr+n}\,\underline{|ns-n}} \cdot r^{nr+n} s^{ns-n}$$
$$= \frac{(nr+n)(nr+n-1)\dots(nr+1)}{ns(ns-1)\dots(ns-n+1)}\,\frac{s^n}{r^n}$$
$$= \frac{(nrs+ns)(nrs+ns-s)\dots(nrs+s)}{(nrs-nr+r)(nrs-nr+2r)\dots nrs},$$
if we write the denominatorial factors in reverse order. Thus we have, dividing the factors by n:

$$x = \frac{rs+s}{rs-r+\frac{r}{n}}\;\frac{rs+s-\frac{s}{n}}{rs-s+\frac{2r}{n}}\;\frac{rs+s-\frac{2s}{n}}{rs-s+\frac{2r}{n}}\dots\frac{rs+\frac{s}{n}}{rs} \quad \dots\dots\dots\dots\dots\text{(v)}.$$

Now all these factor-ratios are greater than unity and the first is the greatest. Their values all lie between

$$\frac{rs+s}{rs-r+\frac{r}{n}} \text{ and } \frac{rs+\frac{s}{n}}{rs}.$$

Bernoulli assumes that the mth factor-ratio will be $\frac{rs+s}{rs}$, and determines m from

$$\frac{rs+s-\frac{m-1}{n}s}{rs-r+\frac{mr}{n}} = \frac{rs+s}{rs} = \frac{r+1}{r} \quad \dots\dots\dots\dots\dots\dots\dots\text{(vi)}.$$

He notes that all the factor-ratios above the mth are greater than the mth and all the following factor-ratios are greater than unity. Thus he finds that

$$x > \left(\frac{r+1}{r}\right)^m.$$

Accordingly if we choose n so that

$$c\,(s-1) \leqslant \left(\frac{r+1}{r}\right)^m \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\text{(vii)},$$

$c\,(s-1)$ will certainly be less than x which is our fundamental condition. But it follows from (vi) and (vii) that

$$n = m\left(1+\frac{s}{r+1}\right) - \frac{s}{r+1},$$
$$m \geqslant \frac{\log c\,(s-1)}{\log(r+1)-\log r}.$$

Similar reasoning applies to the other side of the maximum term. It accordingly follows that if n be given the larger of the two values below, i.e.

$$\left.\begin{array}{l} \dfrac{\log c\,(s-1)}{\log(r+1)-\log r}\left(1+\dfrac{s}{r+1}\right)-\dfrac{s}{r+1} \\[2ex] \dfrac{\log c\,(r-1)}{\log(s+1)-\log s}\left(1+\dfrac{r}{s+1}\right)-\dfrac{r}{s+1} \end{array}\right\} \dots\dots\dots\dots\dots\text{(viii)},$$

then we can make the frequency of the central series of terms larger than c times the frequency of the sum of the tails. But the central series of terms measures the frequency from the $(nr-n)$th term to the $(nr+n)$th term inclusive. Bernoulli now applies this to the theory of probabilities. Let $p_0=r/(r+s)$ be the true probability of success, then in nt trials the probability of a success occurring between $nr-n$ and $nr+n$ times, or an average of successes between $(nr-n)/nt$ and $(nr+n)/nt$, i.e. between $p_0 \pm \frac{1}{t}$, will be measured by the ratio of the sum of the central series of terms to the sum of all the terms. If now we take $\frac{1}{t}$ any small value we please and take n the larger of the two values in (viii), then we are certain that in nt trials the ratio of successes to total trials will have a probability greater than $\frac{c}{c+1}$, or the odds are c to 1 that the ratio of successes to trials will fall within the limits $p_0 \pm \frac{1}{t}$.

Bernoulli then turns the problem round and says that if the observed value in nt trials be p, then the true value p_0 will lie between $p \pm \frac{1}{t}$ with the given probability. This is rather stated than proved, but it is of course the kernel of much later developments of importance. Leibnitz raised objections to it*.

Now let us look at the modern method of dealing with this problem. Suppose p the probability of an event happening, q of its failing, then if the trial be made m times the frequency will be distributed according to the terms of the binomial $(p+q)^m$. The most probable result is mp, and the mean number of successes p. If we want to consider a result which does not differ more than $\pm\,\epsilon$ from p we may obtain its probability by summing the terms of the binomial from $mp-m\epsilon$ to $mp+m\epsilon$, and it can be demonstrated that by taking m sufficiently large this probability can be made as near unity as we please. Provided neither p nor q be indefinitely small, it was first shown by De Moivre and the proof has often been reproduced since, that the appropriate areas of the curve:

$$y=\frac{N}{\sqrt{2\pi}\sqrt{mpq}}\,e^{-\frac{1}{2}\frac{x^2}{mpq}} \quad \text{..........................(ix)},$$

give very approximately the sums of corresponding terms of the binomial. Thus the probability P_ϵ of the observed result lying between $p \pm \epsilon$ is

$$P_\epsilon=\frac{1}{\sqrt{2\pi}\sqrt{mpq}}\int_{-m\epsilon}^{+m\epsilon} e^{-\frac{1}{2}\frac{x^2}{mpq}}\,dx,$$

or:

$$P_\epsilon=\frac{2}{\sqrt{2\pi}}\int_0^{\frac{\sqrt{m}\epsilon}{\sqrt{pq}}} e^{-\frac{1}{2}x'^2}\,dx' \quad \text{..............................(x)}.$$

Accordingly if we indefinitely increase m the upper limit approaches infinity and the value of the integral $\frac{1}{2}\sqrt{2\pi}$ or P_ϵ approaches unity. Or again if we need

---

* See our p. 206.

P_ϵ to have a definite value for a given ϵ, we can find a value of m which will insure it. For a given value of P_ϵ we must have $\sqrt{m}\epsilon/\sqrt{pq}$ a constant, or the limit of ϵ varies inversely as $\sqrt{m}$, i.e. accuracy varies directly as the square root of the number of observations. Such is the so-called Bernoulli's Theorem given by the text-book writers from Cournot in 1840 to the present day. The discussion is essentially that of De Moivre though his name is not mentioned. The French and German writers have consciously or unconsciously taken De Moivre's results and his very method and attributed them to Bernoulli so that practically the whole scientific world believes Bernoulli provided what is alone due to De Moivre. Bernoulli knew nothing of the approach of the normal curve to the high power binomial, he knew nothing of the inverse square root law of accuracy; he obtained only the crudest approximations to the number of observations needful to get a certain value of the odds.

Bernoulli himself gives an example, namely where $r : s :: 3 : 2$. He takes $r = 30$, $s = 20$, $t = 50$ and determines the number of trials which will certainly give odds of $c : 1$ that the observed result will lie between $\frac{3}{5} \pm \cdot 02$. I place here his results against those obtained by De Moivre's Theory:

Odds c : 1	Bernoulli's Result	De Moivre's Result
1000 : 1	25,550 trials	6,498 trials
10,000 : 1	31,258 „	9,082 „
100,000 : 1	36,966 „	11,704 „

These numbers, I think, show that while Bernoulli had a vision of "Bernoulli's Problem," he failed to solve it in anything like a reasonable manner. His discussion is not only lengthy and somewhat obscure, but is too loose to give any valid solution. He appreciated, what most experimenters knew already, that one gets nearer to the truth by increasing the size of the sample, but he really failed to obtain any adequate measure of the approach. Yet Bernoulli himself thought his results of "high utility." The method by which he reached his problem is of much interest. After giving a number of definitions of probability, luck, certitude, moral certitude, etc., Bernoulli states that what he aims at in his theorem is to reach by repeated trials a "moral" certitude of 99 in 100 or 999 in 1000; he wants to show us that repetition will give us a moral certitude of as great an intensity as we please. He fully recognises indeed the importance of the problem under consideration; thus he writes:

> This is therefore the problem that I now wish to publish here, having considered it closely for a period of twenty years, and it is a problem of which the novelty as well as the high utility together with its grave difficulty exceed in value all the remaining chapters of my doctrine. Before I treat of this "Golden Theorem" I will show that a few objections, which certain learned men* have raised against my propositions, are not valid (*Ars Conjectandi*, 1713, p. 327).

---

* See p. 205 above.

Bernoulli admits that as a matter of experience even in his day, the increase of accuracy by the increase of trials was recognised. But he was probably the first to more or less clearly state the problem. It must be admitted, I think, that his twenty years consideration of the matter did not lead him to a solution of much value.

(3) It may not be without interest to deal a little fully with what seem to be the weak points in Bernoulli's method of inequalities. They are essentially two-fold:

(*a*) His method of reaching a value of $x = u_{nr}/u_{nr+n}$ and therefore of n is far too loose.

(*b*) His approximation to the $n(s-2)$ terms in the denominator of (iii) is so appallingly crude that he must get bad results. The idea that the tail may be divided up into $(s-1)$ series of n terms, and that the last $(s-2)$ series may have substituted for their proper values that of the first series, leads of course to an inequality, but one that gives a very poor approximation.

We will consider these points in succession. The actual value of x is given by

$$x = \left(\frac{r}{s}\right)^n \frac{\underline{|nr-n}}{\underline{|nr}} \, \frac{\underline{|ns+n}}{\underline{|ns}}.$$

Now n being large let us apply Stirling's Theorem, and we find

$$x = \sqrt{\frac{r-1}{r}\,\frac{s+1}{s}} \left\{\left(\frac{r-1}{r}\right)^{r-1} \left(\frac{s+1}{s}\right)^{s+1}\right\}^n, \text{ nearly.}$$

Equating this to $c(s-1)$ we find that

$$n \geqslant \frac{\log c(s-1) - \frac{1}{2}\log\frac{r-1}{r} - \frac{1}{2}\log\frac{s+1}{s}}{(r-1)\log\frac{r-1}{r} + (s+1)\log\frac{s+1}{s}} \quad \text{.................(xi).}$$

Had we worked with the series on the other side of the maximum term, we should have had:

$$n \geqslant \frac{\log c(r-1) - \frac{1}{2}\log\frac{s-1}{s} - \frac{1}{2}\log\frac{r+1}{r}}{(s-1)\log\frac{s-1}{s} + (r+1)\log\frac{r+1}{r}} \quad \text{.................(xii).}$$

Computing n from these limits I found for the greater n the values indicated below:

Odds c : 1	Bernoulli	Bernoulli modified by better x	De Moivre
1000 : 1	25,550 trials	12,387 trials	6,498 trials
10,000 : 1	31,258 „	15,165 „	9,082 „
100,000 : 1	36,966 „	17,943 „	11,704 „

We have therefore divided Bernoulli's number of trials roughly by two, when we use a closer approximation than he used to the value of x. We now ask: is it not feasible to obtain a better approximation to the tail of a binomial than Bernoulli adopted? The answer is most certainly and in the simplest of manners. Consider the series beyond the term $nr - n$ in r, i.e.

$$S_{nr-n} = u_{nr-n-1} + u_{nr-n-2} + \ldots\ldots + u_0$$
$$= u_{nr-n}\left(\frac{s}{r}\,\frac{nr-n}{ns+n+1} + \frac{s^2}{r^2}\,\frac{(nr-n)(nr-n-1)}{(ns+n+1)(ns+n+2)} + \ldots\right).$$

Now
$$ns + n + i \text{ is } > ns + n,$$
$$nr - n - i \text{ is } < nr - n;$$

it follows accordingly that

$$S_{nr-n} \text{ is } < u_{nr-n}\left\{\frac{s}{r}\,\frac{nr-n}{ns+n} + \left(\frac{s}{r}\right)^2\left(\frac{nr-n}{ns+n}\right)^2 + \ldots\right.$$
$$< u_{nr-n}\left\{\frac{r-1}{r}\,\frac{s}{s+1} + \left(\frac{r-1}{r}\,\frac{s}{s+1}\right)^2 + \left(\frac{r-1}{r}\,\frac{s}{s+1}\right)^3 + \ldots\right.$$
$$< \frac{u_{nr-n}\dfrac{r-1}{r}\,\dfrac{s}{s+1}}{1 - \dfrac{r-1}{r}\,\dfrac{s}{s+1}} < u_{nr-n}\,\frac{s(r-1)}{t},$$

summing to infinity, which can only increase the right-hand side.

Applying Stirling's Theorem to the term

$$u_{nr-n} = \frac{|\underline{nt}}{|\underline{nr-n}\;|\underline{ns+n}}\,r^{nr-n}\,s^{ns+n},$$

we find:
$$S_{nr-n} < \tfrac{1}{2}t^n \text{ [illegible] } \qquad .(\text{xiii}).$$

where [illegible]

Similarly if S_{ns-n} denote the sum of all the terms with values of the powers of s below $ns - n$, we have:

$$\left.\begin{aligned} S_{ns-n} &< \tfrac{1}{2}t^{nt}\sqrt{\frac{2}{\pi}}\,\frac{1}{\sqrt{n}}\,\beta_2^{\,n}\sqrt{\frac{r^3(s-1)}{t(r+1)}} \\ \text{where}\qquad \beta_2 &= \left(\frac{s}{s-1}\right)^{s-1}\left(\frac{r}{r+1}\right)^{r+1} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots..(\text{xiv}).$$

Now $\frac{1}{2}t^{nt}$ is half the sum of the terms of the total binomial and it is clear that we have obtained a superior limit to the ratio of the tails to the body, or if we please the sum of the tails to the total binomial. The method is at least in one respect superior to De Moivre's, for the latter makes both tails equal, while in the present method we can obtain a limit to either tail independently.

Now following Bernoulli's notation:

$$c = \text{ratio } (\tfrac{1}{2}t^{nt} - \text{tail}) \text{ to tail},$$

or
$$\frac{1}{c+1} = \text{ratio of tail to } \tfrac{1}{2}t^{nt}.$$

Now take $\zeta_1 = 2n \log_e \beta_1$, $\zeta_2 = 2n \log_e \beta_2$, then:

$$\frac{e^{\zeta_1}}{\zeta_1} = \frac{\pi}{4(c+1)^2 \log_e \beta_1} \frac{t(s+1)}{s^2(r-1)}$$

$$= \frac{\cdot 341{,}0940}{(c+1)^2 \log_{10} \beta_1} \frac{t(s+1)}{s^2(r-1)} \quad \text{......(xv)}$$

and

$$\frac{e^{\zeta_2}}{\zeta_2} = \frac{\cdot 341{,}0940}{(c+1)^2 \log_{10} \beta_2} \frac{t(r+1)}{r^2(s-1)} \quad \text{......(xvi)}.$$

If given values now be chosen for c, r, s and t and the above equations solved for ζ_1 and ζ_2 we obtain values of n, and thus of nt, fixing the number of trials needful to obtain given odds c to 1 so that the result will lie within the limits $\frac{r}{r+s} \pm \frac{1}{t}$.

Applying our result to Bernoulli's first case $c = 1000$ we find:

$$\zeta_1 = -10{\cdot}905{,}205, \text{ giving } nt = 6470,$$
$$\zeta_2 = -10{\cdot}898{,}971, \text{ giving } nt = 6502,$$

results of the same order as the 6498 of the normal curve, and approximately again halving the result found by improving x. It is therefore clear that of Bernoulli's exaggerated limit, which is four times De Moivre's value, one half is due to his bad determination of x and one quarter to his poor method of obtaining a limit to the tail. We can now sum up our results in the following table:

Number of Trials needful to obtain the odds given in the first column that with an actual probability of success of ·6, *the observed result will lie between* ·6 ± ·02.

Odds c : 1	Bernoulli's Result	Bernoulli's Result with better choice of x	Tails as Geometrical Series	De Moivre's Value
1000 : 1	25,550 trials	12,387 trials	6,470 6,502 } trials	6,498 trials
10,000 : 1	31,258 „	15,165 „	9,005 9,050 } „	9,082 „
100,000 : 1	36,966 „	17,943 „	11,588 11,647 } „	11,704 „

The reader will be surprised possibly to notice that the geometrical series method gives lower and therefore more accurate results than the normal curve, although the differences are not very important. It has also the advantage of giving different limits for the two tails, which the normal curve does not distinguish owing to its symmetry.

But there is more than one way of applying the normal curve to binomial data. The usual way would be that which gives the numbers in the fifth column above. Namely if $(p+q)^m$ be the binomial, we need the frequency between $p \pm 1/t$, i.e. for the tail we need the area of the normal curve from 0 to $m\frac{1}{t}$. But the area from $-\infty$ to x/σ is $N\frac{1}{2}(1+\alpha)$ in the notation of the *Tables for Statisticians and Biometricians*. In other words $\frac{1}{2}(1+\alpha)$ corresponds to the value $\frac{m\frac{1}{t}}{\sqrt{mpq}}$ of x

in the table of the probability integral. But in Bernoulli's case $t=50$, $1/t=\cdot 02$, $p=\cdot 6$, $q=\cdot 4$ and therefore $\sqrt{m}\times \cdot 02/\sqrt{24}=x$, i.e. $m=600x^2$. But Bernoulli's $c=(\frac{1}{2}\text{body}-\text{tail})/\text{tail}$ or $\frac{1}{2}(1+\alpha)=1-\frac{1}{2(c+1)}=\frac{2c+1}{2c+2}$. It is accordingly easy to find $\frac{1}{2}(1+\alpha)$ and thence x and so m the power of the binomial. I call this the ordinary or usual process, but it overlooks the discrete character of the binomial. We suppose our curve *areas* to represent the discrete values of the terms of the binomial. Now Bernoulli is seeking the sum of the terms from $nr-n$ to $nr+n$, *including both these.* Hence we ought to integrate the curve not from 0 to n but to midway between the nth and $(n+1)$th term, i.e. up to $n+\frac{1}{2}$.

In Bernoulli's notation x/σ will be $\frac{n+\frac{1}{2}}{\sqrt{nt\,.\,\frac{3}{5}\,.\,\frac{2}{5}}}=\frac{n+\frac{1}{2}}{\sqrt{12n}}$, since $t=50$, and this will be the value of the argument, x' say, of the table of the probability integral. Accordingly we have for $\sqrt{n}$

$$\sqrt{n}=\sqrt{3}x'+\sqrt{3x'^2-\cdot 5}\quad\text{.........................(xvii)},$$

x' as before being found from the argument of

$$\tfrac{1}{2}(1+\alpha)=(2c+1)/(2c+2)\quad\text{...................(xviii)}.$$

When $\sqrt{n}$ has been found from this equation, the power m of the binomial is

$$nt=50n.$$

I worked out the values of m for the three cases given by Bernoulli and found them to be

6448, 9032 and 11654.

These make the normal curve method very close indeed in the number of trials to the geometrical series method. The only advantage remaining to the latter is the evaluation of the separate tails. It is a somewhat longer process at present, but this might be remedied by the publication of a table of $e^{-\xi}/\xi$, which could be readily constructed for the suitable range of ξ values.

After all, I think, we must conclude that it is somewhat a perversion of historical facts to call the method involved in Column (v) of the Table on p. 209 by the name of the man who after twenty years of consideration had not got further than the crude values of Column (ii) with their 200 to 300 per cent. excesses. Bernoulli saw the importance of a certain problem; so did Ptolemy, but it would be rather absurd to call Kepler's or Newton's solution of planetary motion by Ptolemy's name! Yet an error of like magnitude seems to me made when De Moivre's method is discussed without reference to its author, under the heading of "Bernoulli's Theorem." The contributions of the Bernoullis to mathematical science are considerable, but they have been in more than one instance greatly exaggerated. The *Pars Quarta* of the *Ars Conjectandi* has not the importance which has often been attributed to it.

I have to thank Mr James Henderson for aid in the interpretation of Bernoulli's somewhat obscure analysis.

ON THE MEANS, STANDARD DEVIATIONS, CORRELATIONS, AND FREQUENCY DISTRIBUTIONS OF FUNCTIONS OF VARIATES.

By Professor KAZUTARO YASUKAWA, M.Sc.

(1) The scope of this paper may be described in the following manner: Two variates x and y have their frequency constants known. u is a single-valued function of x and w a single-valued function of y. It is proposed to study the frequency constants of u and w with their correlations etc., on the assumption that the coefficients of variation of x and y are of the order of magnitude that occurs in the great bulk of biometric investigations. It is not contended that coefficients of variation cannot be found—such as those for weight of human adults or for mortality rates for heterogeneous population-masses—which are unsuited to the expansions herein discussed. But in most anthropometric inquiries coefficients of variation range from 2·0 to 8·0, and accordingly the ratio of standard deviation to mean is from about ·02 to ·08, a reasonable average value being ·04. It is therefore justifiable to expand our functions in series of powers of the coefficients of variation. Expansions to the third or fourth power are generally adequate for statistical purposes, and in some cases it is enough to proceed to the second.

(2) The notation we shall adopt will be of the usual type. For variate x:

$$\bar{x},\ \sigma_x,\ v_x = \sigma_x/\bar{x},\quad {}_x\beta_1 = {}_x\mu_3^2/{}_x\mu_2^3,\quad {}_x\beta_2 = {}_x\mu_4/{}_x\mu_2^2,$$

and for variate y:

$$\bar{y},\ \sigma_y,\ v_y = \sigma_y/\bar{y},\quad {}_y\beta_1 = {}_y\mu_3^2/{}_y\mu_2^3,\quad {}_y\beta_2 = {}_y\mu_4/{}_y\mu_2^2,$$

while r_{xy}, r_{uw} will represent correlations; $\bar{u}$, $\bar{w}$ the means of u and w respectively functions of x and y, and σ_u, σ_w their standard deviations. When a variate is dashed, it signifies that it is measured from its mean.

Now $u = \phi_1(x)$ may be written

$$\bar{u} + u' = \phi_1(\bar{x} + x') = f_1\left(1 + \frac{x'}{\bar{x}}\right),$$

and it is this function f_1, or comparable functions, that we shall use throughout.

Expanding $f_1(1 + x'/\bar{x})$ by Taylor's theorem we obtain

$$u + u' = f_1(1) + \frac{x'}{\bar{x}} f_1'(1) + \tfrac{1}{2}\left(\frac{x'}{\bar{x}}\right)^2 f_1''(1) + \tfrac{1}{6}\left(\frac{x'}{\bar{x}}\right)^3 f_1'''(1)$$

$$+ \tfrac{1}{24}\left(\frac{x'}{\bar{x}}\right)^4 f_1^{\text{iv}}(1) + \text{etc.} \quad\ldots\ldots\ldots(1).$$

Summing for all values of x' and dividing by their number, we have:

$$\bar{u} = f_1(1) + \tfrac{1}{2} v_x^2 f_1''(1) + \tfrac{1}{6} \sqrt{{}_x\beta_1}\, v_x^3 f_1'''(1) + \tfrac{1}{24}\, {}_x\beta_2 v_x^4 f_1^{\text{iv}}(1) + \dots \quad \dots(2).$$

This series which expresses the mean value of a function of x in terms of the frequency constants of x will converge, owing not only to the customary smallness of v_x, but also to the usual convergency of the terms of a Taylor's series.

Subtracting (2) from (1) we find:

$$u' = \frac{x'}{\bar{x}} f_1'(1) + \tfrac{1}{2} f_1''(1) \left\{ \left(\frac{x'}{\bar{x}}\right)^2 - v_x^2 \right\} + \tfrac{1}{6} f_1'''(1) \left\{ \left(\frac{x'}{\bar{x}}\right)^3 - \sqrt{{}_x\beta_1}\, v_x^3 \right\}$$
$$+ \tfrac{1}{24} f_1^{\text{iv}}(1) \left\{ \left(\frac{x'}{\bar{x}}\right)^4 - {}_x\beta_2 v_x^4 \right\} + \dots \quad \dots(3),$$

or, introducing v_x and σ_x and rearranging:

$$u' = v_x f_1'(1) \left[\frac{x'}{\sigma_x} + \tfrac{1}{2} v_x \frac{f_1''(1)}{f_1'(1)} \left\{ \left(\frac{x'}{\sigma_x}\right)^2 - 1 \right\} + \tfrac{1}{6} v_x^2 \frac{f_1'''(1)}{f_1'(1)} \left\{ \left(\frac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1} \right\} \right.$$
$$\left. + \tfrac{1}{24} v_x^3 \frac{f_1^{\text{iv}}(1)}{f_1'(1)} \left\{ \left(\frac{x'}{\sigma_x}\right)^4 - {}_x\beta_2 \right\} + \dots \right] \dots\dots\dots(4).$$

Squaring (4) and taking mean values we reach after some troublesome but straightforward algebra:

$$\sigma_u^2 = v_x^2 (f_1'(1))^2 \left[1 + v_x \frac{f_1''(1)}{f_1'(1)} \sqrt{{}_x\beta_1} + \tfrac{1}{3} v_x^2 \frac{f_1'''(1)}{f_1'(1)}\, {}_x\beta_2 \right.$$
$$+ \tfrac{1}{4} v_x^2 \left(\frac{f_1''(1)}{f_1'(1)}\right)^2 ({}_x\beta_2 - 1) + \tfrac{1}{12} v_x^3 \frac{f_1^{\text{iv}}(1)}{f_1'(1)}\, {}_x\beta_3'$$
$$\left. + \tfrac{1}{6} v_x^3 \frac{f_1''(1) f_1'''(1)}{(f_1'(1))^2} ({}_x\beta_3' - \sqrt{{}_x\beta_1}) + \dots \right] \dots\dots\dots\dots\dots\dots(5),$$

where $${}_x\beta_3' = {}_x\mu_5/\sigma_x^5 = {}_x\beta_3/\sqrt{{}_x\beta_1}.$$

Taking the square root we have:

$$\sigma_u = \pm\, v_x f_1'(1) \left[1 + \tfrac{1}{2} v_x \frac{f_1''(1)}{f_1'(1)} \sqrt{{}_x\beta_1} + \tfrac{1}{6} v_x^2 \frac{f_1'''(1)}{f'(1)}\, {}_x\beta_2 \right.$$
$$+ \tfrac{1}{8} v_x^2 \left(\frac{f_1''(1)}{f_1'(1)}\right)^2 ({}_x\beta_2 - {}_x\beta_1 - 1) + \tfrac{1}{24} v_x^3 \frac{f_1^{\text{iv}}(1)}{f_1'(1)}\, {}_x\beta_3'$$
$$+ \tfrac{1}{12} v_x^3 \frac{f_1''(1) f_1'''(1)}{(f_1'(1))^2} ({}_x\beta_3' - {}_x\beta_2 \sqrt{{}_x\beta_1} - \sqrt{{}_x\beta_1})$$
$$\left. - \tfrac{1}{16} v_x^3 \left(\frac{f_1''(1)}{f_1'(1)}\right)^3 ({}_x\beta_2 - {}_x\beta_1 - 1) \sqrt{{}_x\beta_1} + \dots \right] \dots\dots\dots\dots\dots(6).$$

Of the above double sign we must choose the one which always makes σ_u positive, i.e. the same sign as that of $f_1'(1)$.

This is the value up to terms of the fourth order in v_x of the standard deviation of the function of a variate.

Dividing (4) by (6):

$$\frac{u'}{\sigma_u}=\pm^{*}\frac{\left[\frac{x'}{\sigma_x}+\frac{1}{2}v_x\frac{f_1''(1)}{f_1'(1)}\left\{\left(\frac{x'}{\sigma_x}\right)^2-1\right\}+\frac{1}{6}v_x^2\frac{f_1'''(1)}{f_1'(1)}\left\{\left(\frac{x'}{\sigma_x}\right)^3-\sqrt{{}_x\beta_1}\right\}+\frac{1}{24}v_x^3\frac{f_1^{\text{iv}}(1)}{f_1'(1)}\left\{\left(\frac{x'}{\sigma_x}\right)^4-{}_x\beta_2\right\}+\dots\right]}{\left[\begin{array}{l}1+\frac{1}{2}v_x\frac{f_1''(1)}{f_1'(1)}\sqrt{{}_x\beta_1}+\frac{1}{6}v_x^2\frac{f_1'''(1)}{f_1'(1)}{}_x\beta_2+\frac{1}{8}v_x^2\left(\frac{f_1''(1)}{f_1'(1)}\right)^2({}_x\beta_2-{}_x\beta_1-1)\\ \quad+\frac{1}{2}v_x^3\frac{f_1^{\text{iv}}(1)}{f_1'(1)}{}_x\beta_3'+\frac{1}{12}v_x^3\frac{f_1''(1)f_1'''(1)}{(f_1'(1))^2}({}_x\beta_3'-{}_x\beta_2\sqrt{{}_x\beta_1}-\sqrt{{}_x\beta_1})\\ \quad-\frac{1}{16}v_x^3\left(\frac{f_1''(1)}{f_1'(1)}\right)^3({}_x\beta_2-{}_x\beta_1-1)\sqrt{{}_x\beta_1}+\dots\end{array}\right.}\qquad\dots\dots(7).$$

Multiply both sides of (7) by $\frac{x'}{\sigma_x}$ and take the mean of this product:

$$r_{ux}=\pm^{*}\left[1-\frac{1}{8}v_x^2\left(\frac{f_1''(1)}{f_1'(1)}\right)^2({}_x\beta_2-{}_x\beta_1-1)-\frac{1}{12}v_x^3\frac{f_1''(1)f_1'''(1)}{(f_1'(1))^2}({}_x\beta_3'-{}_x\beta_2\sqrt{{}_x\beta_1}-\sqrt{{}_x\beta_1})\right.$$
$$\left.+\frac{1}{8}v_x^3\left(\frac{f_1''(1)}{f_1'(1)}\right)^3\sqrt{{}_x\beta_1}({}_x\beta_2-{}_x\beta_1-1)+\dots\right]\quad\dots\dots(8).$$

This is the correlation coefficient between a variate and any function of this variate, and indicates how far the coefficient diverges numerically from unity. The true association between the variates would be measured by the correlation ratio, $\eta_{u.x}=1$. The terms in v_x^2 etc. represent how far the association as measured by the coefficient of correlation falls short of the correlation ratio.

If the variate x follows the normal law,

$${}_x\beta_1={}_x\beta_3'=0,\quad {}_x\beta_2=3,$$

and $$r_{ux}=\pm^{\dagger}\left[1-\frac{1}{4}v_x^2\left(\frac{f_1''(1)}{f_1'(1)}\right)^2+\text{terms in } v_x^4 \text{ and higher powers}\right]\quad\dots(9).$$

Table I provides a series of values of r_{ux} for various functions u of x, when we give different values to v_x. The reader may convince himself from this table how closely the values of the correlation coefficient approach unity—for all practical

TABLE I.

Approximate Values of r_{ux} for various Functions of x and various Values of v_x.

Values of v_x.

Form of u	Approximate value of r_{ux}	*·01*	*·02*	*·03*	*·04*	*·05*	*·06*	*·07*	*·08*	*·09*	*·10*	*·12*
$u=kx^3;\quad\frac{f_1''(1)}{f_1'(1)}=2$	$1-v_x^2$	·9999	·9996	·9991	·9984	·9975	·9964	·9951	·9936	·9919	·9900	·9856
$u=\frac{k}{x};\quad\frac{f_1''(1)}{f_1'(1)}=-2$	$1-v_x^2$	·9999	·9996	·9991	·9984	·9975	·9964	·9951	·9936	·9919	·9900	·9856
$u=k\sqrt{x};\quad\frac{f_1''(1)}{f_1'(1)}=-\frac{1}{2}$	$1-\left(\frac{v_x}{4}\right)^2$	1·0000	1·0000	·9999	·9999	·9998	·9998	·9997	·9996	·9995	·9994	·9991
$u=k\log x;\quad\frac{f_1''(1)}{f_1'(1)}=-1$	$1-\left(\frac{v_x}{2}\right)^2$	1·0000	·9999	·9998	·9996	·9994	·9991	·9988	·9984	·9980	·9975	·9964

---

* The sign chosen in (7) and (8) must be the same as that adopted for (6) as stated in the text.

† The sign to be selected again as in (6).

purposes—even with considerable values of v_x. The practical statistician dealing with series of the customary size would have little hesitation in asserting a causal relation to exist between u and x.

We will now suppose w is a function of the second variate y, or

$$w = \phi_2(y) = f_2\left(1 + \frac{y'}{y}\right).$$

We have expressions for $\bar{w}$, w' and σ_w exactly analogous to those for $\bar{u}$, u' and σ_u in (2), (3) and (6), except that f_2 now replaces f_1, v_y replaces v_x and the β's of y those of x. In order to abbreviate the results we will use single symbols for the functional constants, namely

$$\lambda_1'' = \frac{f_1''(1)}{f_1'(1)}, \quad \lambda_1''' = \frac{f_1'''(1)}{f_1'(1)}, \quad \lambda_1^{\text{iv}} = \frac{f_1^{\text{iv}}(1)}{f_1'(1)},$$
$$\lambda_2'' = \frac{f_2''(1)}{f_2'(1)}, \quad \lambda_2''' = \frac{f_2'''(1)}{f_2'(1)}, \quad \lambda_2^{\text{iv}} = \frac{f_2^{\text{iv}}(1)}{f_2'(1)}.$$

We may now write down

$$\frac{u'}{\sigma_u} = \pm^{*}\frac{\dfrac{x'}{\sigma_x} + \frac{1}{2}v_x\lambda_1''\left\{\left(\dfrac{x'}{\sigma_x}\right)^2 - 1\right\} + \frac{1}{6}v_x^2\lambda_1'''\left\{\left(\dfrac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1}\right\} + \frac{1}{24}v_x^3\lambda_1^{\text{iv}}\left\{\left(\dfrac{x'}{\sigma_x}\right)^4 - {}_x\beta_2\right\}}{\left[\begin{array}{c}1 + \frac{1}{2}v_x\lambda_1''\sqrt{{}_x\beta_1} + \frac{1}{6}v_x^2\lambda_1'''\,{}_x\beta_2 + \frac{1}{8}v_x^2\lambda_1''^2({}_x\beta_2 - {}_x\beta_1 - 1) + \frac{1}{12}v_x^3\lambda_1''\lambda_1'''({}_x\beta_3' - {}_x\beta_2\sqrt{{}_x\beta_1} - \sqrt{{}_x\beta_1})\\ + \frac{1}{24}v_x^3\lambda_1^{\text{iv}}\,{}_x\beta_3' - \frac{1}{16}v_x^3\lambda_1''^3({}_x\beta_2 - {}_x\beta_1 - 1)\sqrt{{}_x\beta_1}\end{array}\right]}$$

$$\equiv \frac{N_1}{D_1} \text{ say} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (10),$$

$$\frac{w'}{\sigma_w} = \pm^{*}\frac{\dfrac{y'}{\sigma_y} + \frac{1}{2}v_y\lambda_2''\left\{\left(\dfrac{y'}{\sigma_y}\right)^2 - 1\right\} + \frac{1}{6}v_y^2\lambda_2'''\left\{\left(\dfrac{y'}{\sigma_y}\right)^3 - \sqrt{{}_y\beta_1}\right\} + \frac{1}{24}v_y^3\lambda_2^{\text{iv}}\left\{\left(\dfrac{y'}{\sigma_y}\right)^4 - {}_y\beta_2\right\}}{\left[\begin{array}{c}1 + \frac{1}{2}v_y\lambda_2''\sqrt{{}_y\beta_1} + \frac{1}{6}v_y^2\lambda_2'''\,{}_y\beta_2 + \frac{1}{8}v_y^2\lambda_2''^2({}_y\beta_2 - {}_y\beta_1 - 1) + \frac{1}{12}v_y^3\lambda_2''\lambda_2'''({}_y\beta_3' - {}_y\beta_2\sqrt{{}_y\beta_1} - \\ + \frac{1}{24}v_y^3\lambda_2^{\text{iv}}\,{}_y\beta_3' - \frac{1}{16}v_y^3\lambda_2''^3({}_y\beta_2 - {}_y\beta_1 - 1)\sqrt{{}_y\beta_1}\end{array}\right]}$$

$$\equiv \frac{N_2}{D_2} \text{ say} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (11).$$

Multiplying (10) and (11) and taking the mean we get for the mean numerator:

$$\begin{aligned}
&\frac{1}{N}S\left[\frac{x'}{\sigma_x} + \tfrac{1}{2}v_x\lambda_1''\left\{\left(\frac{x'}{\sigma_x}\right)^2 - 1\right\} + \tfrac{1}{6}v_x^2\lambda_1'''\left\{\left(\frac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1}\right\} + \tfrac{1}{24}v_x^3\lambda_1^{\text{iv}}\left\{\left(\frac{x'}{\sigma_x}\right)^4 - {}_x\beta_2\right\}\right]\\
&\times\left[\frac{y'}{\sigma_y} + \tfrac{1}{2}v_y\lambda_2''\left\{\left(\frac{y'}{\sigma_y}\right)^2 - 1\right\} + \tfrac{1}{6}v_y^2\lambda_2'''\left\{\left(\frac{y'}{\sigma_y}\right)^3 - \sqrt{{}_y\beta_1}\right\} + \tfrac{1}{24}v_y^3\lambda_2^{\text{iv}}\left\{\left(\frac{y'}{\sigma_y}\right)^4 - {}_y\beta_2\right\}\right]\\
&= \frac{1}{N}S\left[\frac{x'y'}{\sigma_x\sigma_y} + \tfrac{1}{2}v_x\lambda_1''\left\{\frac{y'}{\sigma_y}\left(\frac{x'}{\sigma_x}\right)^2 - \frac{y'}{\sigma_y}\right\} + \tfrac{1}{2}v_y\lambda_2''\left\{\frac{x'}{\sigma_x}\left(\frac{y'}{\sigma_y}\right)^2 - \frac{x'}{\sigma_x}\right\}\right.\\
&\quad + \tfrac{1}{6}v_x^2\lambda_1'''\left\{\frac{y'}{\sigma_y}\left(\frac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1}\,\frac{y'}{\sigma_y}\right\} + \tfrac{1}{6}v_y^2\lambda_2'''\left\{\frac{x'}{\sigma_x}\left(\frac{y'}{\sigma_y}\right)^3 - \sqrt{{}_y\beta_1}\,\frac{x'}{\sigma_x}\right\}\\
&\quad + \tfrac{1}{4}v_xv_y\lambda_1''\lambda_2''\left\{\frac{S(x'y')^2}{\sigma_x^2\sigma_y^2} - \frac{x'^2}{\sigma_x^2} - \frac{y'^2}{\sigma_y^2} + 1\right\}\\
&\quad + \tfrac{1}{24}v_x^3\lambda_1^{\text{iv}}\left\{\frac{y'}{\sigma_y}\left(\frac{x'}{\sigma_x}\right)^4 - {}_x\beta_2\,\frac{y'}{\sigma_y}\right\} + \tfrac{1}{24}v_y^3\lambda_2^{\text{iv}}\left\{\frac{x'}{\sigma_x}\left(\frac{y'}{\sigma_y}\right)^4 - {}_y\beta_2\,\frac{x'}{\sigma_x}\right\}\\
&\quad + \tfrac{1}{12}v_x^2v_y\lambda_1'''\lambda_2''\left\{\frac{(x'^3y'^2)}{\sigma_x^3\sigma_y^2} - \sqrt{{}_x\beta_1}\,\frac{y'^2}{\sigma_y^2} - \frac{x'^3}{\sigma_x^3} + \sqrt{{}_x\beta_1}\right\}\\
&\quad \left.+ \tfrac{1}{12}v_xv_y^2\lambda_1''\lambda_2'''\left\{\frac{(x'^2y'^3)}{\sigma_x^2\sigma_y^3} - \sqrt{{}_y\beta_1}\,\frac{x'^2}{\sigma_x^2} - \frac{y'^3}{\sigma_y^3} + \sqrt{{}_y\beta_1}\right\}\right].
\end{aligned}$$

---

* The sign to be chosen as before.

Now in this expression nine mean products occur:

$$\frac{1}{N}\frac{S(y'x'^2)}{\sigma_y\sigma_x^2},\ \frac{1}{N}\frac{S(x'y'^2)}{\sigma_x\sigma_y^2},\ \frac{1}{N}\frac{S(x'^2y'^2)}{\sigma_x^2\sigma_y^2},\ \frac{1}{N}\frac{S(x'y'^3)}{\sigma_x\sigma_y^3},\ \frac{1}{N}\frac{S(y'x'^3)}{\sigma_y\sigma_x^3},$$

$$\frac{1}{N}\frac{S(x'^4y')}{\sigma_x^4\sigma_y},\ \frac{1}{N}\frac{S(x'y'^4)}{\sigma_x\sigma_y^4},\ \frac{1}{N}\frac{S(x'^3y'^2)}{\sigma_x^3\sigma_y^2}\ \text{and}\ \frac{1}{N}\frac{S(x'^2y'^3)}{\sigma_x^2\sigma_y^3}.$$

These are multiplied respectively by $v_x\lambda_1''$, $v_y\lambda_2''$, $v_xv_y\lambda_1''\lambda_2''$, $v_x^2\lambda_1'''$, $v_y^2\lambda_2'''$, $v_x^3\lambda_1^{\text{iv}}$, $v_y^3\lambda_2^{\text{iv}}$, $v_x^2v_y\lambda_1'''\lambda_2''$ and $v_xv_y^2\lambda_1''\lambda_2'''$. The last four are certainly of a fairly low order, the first five of more importance. But it is clear that we cannot without a knowledge of the x and y frequency distributions proceed further unless we make some assumptions, which will provide at any rate approximations to these terms*.

If we assume linear regression then:

$$\frac{1}{N}\frac{S(x'y'^2)}{\sigma_x\sigma_y^2}=r_{xy}\sqrt{{}_y\beta_1},\quad \frac{1}{N}\frac{S(y'x'^2)}{\sigma_y\sigma_x^2}=r_{xy}\sqrt{{}_x\beta_1},$$

$$\frac{1}{N}\frac{S(x'y'^4)}{\sigma_x\sigma_y^4}=r_{xy}\,{}_y\beta_3',\quad \frac{1}{N}\frac{S(y'x'^4)}{\sigma_y\sigma_x^4}=r_{xy}\,{}_x\beta_3',$$

$$\frac{1}{N}\frac{S(x'y'^3)}{\sigma_x\sigma_y^3}=r_{xy}\,{}_y\beta_2,\quad \frac{1}{N}\frac{S(y'x'^3)}{\sigma_y\sigma_x^3}=r_{xy}\,{}_x\beta_2.$$

If we assume in addition homoscedasticity:

$$\frac{1}{N}\frac{S(x'^2y'^2)}{\sigma_x^2\sigma_y^2}=1+\tfrac{1}{2}r_{xy}^2({}_x\beta_2+{}_y\beta_2-2),$$

$$\frac{1}{N}\frac{S(x'^2y'^3)}{\sigma_x^2\sigma_y^3}=(1-r_{xy}^2)\sqrt{{}_y\beta_1}+r_{xy}^2\,{}_y\beta_3',$$

$$\frac{1}{N}\frac{S(y'^2x'^3)}{\sigma_y^2\sigma_x^3}=(1-r_{xy}^2)\sqrt{{}_x\beta_1}+r_{xy}^2\,{}_x\beta_3'.$$

The first four and the last two are all zero for normal distributions of x and y, and accordingly will be small for moderate skewness. The seventh one for normality becomes $1+2r_{xy}^2$, and the fifth and sixth $3r_{xy}$.

With the above results we obtain for the mean values of N_1N_2 and D_1D_2:

$$\begin{aligned}\text{Mean}\,(N_1N_2)=r_{xy}&+\tfrac{1}{2}v_x\lambda_1''\,r_{xy}\sqrt{{}_x\beta_1}+\tfrac{1}{2}v_y\lambda_2''\,r_{xy}\sqrt{{}_y\beta_1}\\&+\tfrac{1}{6}v_x^2\lambda_1'''\,r_{xy}\,{}_x\beta_2+\tfrac{1}{6}v_y^2\lambda_2'''\,r_{xy}\,{}_y\beta_2+\tfrac{1}{8}v_xv_y\lambda_1''\lambda_2''r_{xy}^2({}_x\beta_2+{}_y\beta_2-2)\\&+\tfrac{1}{24}v_x^3\lambda_1^{\text{iv}}\,r_{xy}\,{}_x\beta_3'+\tfrac{1}{24}v_y^3\lambda_2^{\text{iv}}\,r_{xy}\,{}_y\beta_3'\\&+\tfrac{1}{12}v_x^2v_y\lambda_1'''\lambda_2''r_{xy}^2({}_x\beta_3'-\sqrt{{}_x\beta_1})+\tfrac{1}{12}v_xv_y^2\lambda_1''\lambda_2'''r_{xy}^2({}_y\beta_3'-\sqrt{{}_y\beta_1}).\end{aligned}$$

$$\begin{aligned}D_1D_2=1&+\tfrac{1}{2}v_x\lambda_1''\sqrt{{}_x\beta_1}+\tfrac{1}{2}v_y\lambda_2''\sqrt{{}_y\beta_1}+\tfrac{1}{6}v_x^2\lambda_1'''\,{}_x\beta_2+\tfrac{1}{8}v_x^2\lambda_1''^2({}_x\beta_2-{}_x\beta_1-1)\\&+\tfrac{1}{4}v_xv_y\lambda_1''\lambda_2''\sqrt{{}_x\beta_1\,{}_y\beta_1}+\tfrac{1}{6}v_y^2\lambda_2'''\,{}_y\beta_2+\tfrac{1}{8}v_y^2\lambda_2''^2({}_y\beta_2-{}_y\beta_1-1)\\&+\tfrac{1}{12}v_x^3\lambda_1''\lambda_1'''({}_x\beta_3'-{}_x\beta_2\sqrt{{}_x\beta_1}-\sqrt{{}_x\beta_1})+\tfrac{1}{24}v_x^3\lambda_1^{\text{iv}}\,{}_x\beta_3'-\tfrac{1}{16}v_x^3\lambda_1''^3({}_x\beta_2-{}_x\beta_1-1)\sqrt{{}_x\beta_1}\\&+\tfrac{1}{12}v_y^3\lambda_2''\lambda_2'''({}_y\beta_3'-{}_y\beta_2\sqrt{{}_y\beta_1}-\sqrt{{}_y\beta_1})+\tfrac{1}{24}v_y^3\lambda_2^{\text{iv}}\,{}_y\beta_3'-\tfrac{1}{16}v_y^3\lambda_2''^3({}_y\beta_2-{}_y\beta_1-1)\sqrt{{}_y\beta_1}\\&+\tfrac{1}{16}v_x^2v_y\lambda_1''^2\lambda_2''({}_x\beta_2-{}_x\beta_1-1)\sqrt{{}_y\beta_1}+\tfrac{1}{12}v_x^2v_y\lambda_1'''\lambda_2''\sqrt{{}_y\beta_1}\,{}_x\beta_2\\&+\tfrac{1}{16}v_xv_y^2\lambda_1''\lambda_2''^2\sqrt{{}_x\beta_1}({}_y\beta_2-{}_y\beta_1-1)+\tfrac{1}{12}v_xv_y^2\lambda_1''\lambda_2'''\sqrt{{}_x\beta_1}\,{}_y\beta_2.\end{aligned}$$

---

* See *Biometrika*, Vol. IX. pp. 4, 5, 7.

Hence:

$$r_{uw} = \frac{1}{N} S \frac{u'w'}{\sigma_u \sigma_w}$$

$$= \pm {}^{*} \frac{\left[\begin{aligned} & r_{xy} + \tfrac{1}{2} v_x \lambda_1'' r_{xy} \sqrt{{}_x\beta_1} + \tfrac{1}{2} v_y \lambda_2'' r_{xy} \sqrt{{}_y\beta_1} + \tfrac{1}{6} v_x^2 \lambda_1''' r_{xy}\, {}_x\beta_2 + \tfrac{1}{6} v_y^2 \lambda_2''' r_{xy}\, {}_y\beta_2 \\ & + \tfrac{1}{8} v_x v_y \lambda_1'' \lambda_2'' r_{xy}^2 ({}_x\beta_2 + {}_y\beta_2 - 2) + \tfrac{1}{24} v_x^3 \lambda_1^{\text{iv}} r_{xy}\, {}_x\beta_3' + \tfrac{1}{24} v_y^3 \lambda_2^{\text{iv}} r_{xy}\, {}_y\beta_3' \\ & + \tfrac{1}{12} v_x^2 v_y \lambda_1''' \lambda_2'' r_{xy}^2 ({}_x\beta_3' - \sqrt{{}_x\beta_1}) + \tfrac{1}{12} v_y^2 v_x \lambda_1'' \lambda_2''' r_{xy}^2 ({}_y\beta_3' - \sqrt{{}_y\beta_1}) \end{aligned}\right]}{D_1 D_2}.$$

Expanding $D_1 D_2$ we find finally:

$$\begin{aligned} r_{uw} = \pm {}^{\dagger} r_{xy} \Big[& 1 - \tfrac{1}{8} v_x^2 \lambda_1''^2 ({}_x\beta_2 - {}_x\beta_1 - 1) - \tfrac{1}{8} v_y^2 \lambda_2''^2 ({}_y\beta_2 - {}_y\beta_1 - 1) \\ & + \tfrac{1}{8} v_x v_y \lambda_1'' \lambda_2'' \{ r_{xy} ({}_x\beta_2 + {}_y\beta_2 - 2) - 4 \sqrt{{}_x\beta_1 {}_y\beta_1} \} \\ & - \tfrac{1}{12} v_x^3 \lambda_1'' \lambda_1''' ({}_x\beta_3' - {}_x\beta_2 \sqrt{{}_x\beta_1} - \sqrt{{}_x\beta_1}) - \tfrac{1}{12} v_y^3 \lambda_2'' \lambda_2''' ({}_y\beta_3' - {}_y\beta_2 \sqrt{{}_y\beta_1} - \sqrt{{}_y\beta_1}) \\ & + \tfrac{1}{8} v_x^3 \lambda_1''^3 \sqrt{{}_x\beta_1} ({}_x\beta_2 - {}_x\beta_1 - 1) + \tfrac{1}{8} v_y^3 \lambda_2''^3 \sqrt{{}_y\beta_1} ({}_y\beta_2 - {}_y\beta_1 - 1) \\ & + \tfrac{1}{16} v_x^2 v_y \lambda_1''^2 \lambda_2'' \sqrt{{}_x\beta_1} \{ 2 \sqrt{{}_x\beta_1 {}_y\beta_1} - r_{xy} ({}_x\beta_2 + {}_y\beta_2 - 2) \} \\ & + \tfrac{1}{16} v_x v_y^2 \lambda_1'' \lambda_2''^2 \sqrt{{}_y\beta_1} \{ 2 \sqrt{{}_x\beta_1 {}_y\beta_1} - r_{xy} ({}_x\beta_2 + {}_y\beta_2 - 2) \} \\ & - \tfrac{1}{12} v_x^2 v_y \lambda_1''' \lambda_2'' \{ {}_x\beta_2 \sqrt{{}_y\beta_1} + r_{xy} (\sqrt{{}_x\beta_1} - {}_x\beta_3') \} \\ & - \tfrac{1}{12} v_x v_y^2 \lambda_1'' \lambda_2''' \{ {}_y\beta_2 \sqrt{{}_x\beta_1} + r_{xy} (\sqrt{{}_y\beta_1} - {}_y\beta_3') \} \Big] \quad \ldots\ldots\ldots\ldots (12). \end{aligned}$$

For the particular case of x and y having normal frequency:

$$r_{uw} = \pm {}^{\ddagger} r_{xy} [1 - \tfrac{1}{4} v_x^2 \lambda_1''^2 - \tfrac{1}{4} v_y^2 \lambda_2''^2 + \tfrac{1}{2} v_x v_y \lambda_1'' \lambda_2'' r_{xy} + \text{fourth order terms}] \ldots (13).$$

For this special case we easily deduce

$$r_{uw} = \pm {}^{\ddagger} r_{xy} \{ 1 - (\sqrt{1 - r_{ux}} - \sqrt{1 - r_{wy}})^2 \} \ldots\ldots\ldots\ldots (14),$$

and this probably gives in most cases a reasonable approximation.

Table II (p. 217) indicates the magnitude of the percentage changes in the correlation of x and y which we obtain when we correlate functions of x and y instead of x and y; they are calculated on an average value of the coefficients of variation. It will be seen that the percentages reached are inside the probable errors of most correlation coefficients. Even if we raise the coefficients of variation to 8, or the order of v_x and v_y to ·08, we shall then only have to multiply these percentages by four, and we shall still find they are of very small importance. It is not therefore reasonable to anticipate that in the great majority of cases any substantial increase of correlation will be obtained by correlating functions of variates instead of variates themselves.

Actual Numerical Illustration. It appeared to me desirable to illustrate these theoretical conclusions by an actual sample population. I accordingly tabled 1112 cases of stature in Father and Son from Pearson's family measurements. The correlation table is provided in Table III (p. 218). I then took the cubes of the statures as provided in Table IV (p. 219), using as my unit 1000 cubic inches.

[*] In the above expression the positive or negative sign is to be taken according as $f_1'(1) f_2'(1)$ is positive or negative.

[†] Choice of sign must be the same as in the previous equation.

[‡] The choice of sign must as before depend on that of $f_1'(1) f_2'(1)$.

TABLE II. *Correlation of u and w Functions of x and y respectively.*

Formula: $r_{uw} = \pm {}^{*}r_{xy}\left[1 - \frac{1}{4}v_x^2\lambda_1''^2 - \frac{1}{4}v_y^2\lambda_2''^2 + \frac{1}{2}v_x v_y \lambda_1''\lambda_2'' r_{xy}\right]$, with $v_x = v_y = {\cdot}04$.

Values of r_{uw} and $100\,\Delta/r_{xy}$.

Values of the functions u and w		r_{xy}										
		0	·1	·2	·3	·4	·5	·6	·7	·8	·9	1·0
(1) $u=kx^3$	* $\pm r_{uw}$	0	·099712	·199488	·299328	·399232	·499200	·599232	·699328	·799488	·899712	1·0000
$w=k'y^3$	$\frac{100\Delta}{r_{xy}}$	—	·288	·256	·224	·192	·160	·128	·096	·064	·032	0
(2) $u=k\sqrt{x}$	* $\pm r_{uw}$	0	·099982	·199968	·299958	·399952	·499950	·599952	·699958	·799968	·899982	1·0000
$w=k'\sqrt{y}$	$\frac{100\Delta}{r_{xy}}$	—	·018	·016	·014	·012	·010	·008	·006	·004	·002	0
(3) $u=\frac{k}{x}$	* $\pm r_{uw}$	0	·099712	·199488	·299328	·399232	·499200	·599232	·699328	·799488	·899712	1·0000
$w=\frac{k'}{y}$	$\frac{100\Delta}{r_{xy}}$	—	·288	·256	·224	·192	·160	·128	·096	·064	·032	0
(4) $u=k\log x$	* $\pm r_{uw}$	0	·099928	·199872	·299832	·399808	·499800	·599808	·699832	·799872	·899928	1·0000
$w=k'\log y$	$\frac{100\Delta}{r_{xy}}$	—	·072	·064	·056	·048	·040	·032	·024	·016	·008	0
(5) $u=k\sqrt{x}$	* $\pm r_{uw}$	0	·099822	·199628	·299418	·399192	·498950	·598692	·698418	·798128	·897822	·99750
$w=k'y^3$	$\frac{100\Delta}{r_{xy}}$	—	·178	·186	·194	·202	·210	·218	·226	·234	·242	·250
(6) $u=\frac{k}{x}$	* $\pm r_{uw}$	0	·099816	·199664	·299544	·399456	·499400	·599376	·699384	·799424	·899496	·99960
$w=k'\log y$	$\frac{100\Delta}{r_{xy}}$	—	·184	·168	·152	·136	·120	·104	·088	·072	·056	·040
(7) $u=kx^3$	* $\pm r_{uw}$	0	·099784	·199536	·299256	·398944	·498600	·598224	·697816	·797376	·896904	·99640
$w=k'\log y$	$\frac{100\Delta}{r_{xy}}$	—	·216	·232	·248	·264	·280	·296	·312	·328	·344	·360
(8) $u=kx^3$	* $\pm r_{uw}$	0	·099648	·199232	·298752	·398208	·497600	·596928	·696192	·795392	·894528	·99360
$w=\frac{k'}{y}$	$\frac{100\Delta}{r_{xy}}$	—	·352	·384	·416	·448	·480	·512	·544	·576	·608	·640
(9) $u=k\sqrt{x}$	* $\pm r_{uw}$	0	·099838	·199692	·299562	·399448	·499350	·599268	·699202	·799152	·899118	·99910
$w=\frac{k'}{y}$	$\frac{100\Delta}{r_{xy}}$	—	·162	·154	·146	·138	·130	·122	·114	·106	·098	·090
(10) $u=k\sqrt{x}$	* $\pm r_{uw}$	0	·099954	·199916	·299886	·399864	·499850	·599844	·699846	·799856	·899874	·9999
$w=k'\log y$	$\frac{100\Delta}{r_{xy}}$	—	·046	·042	·038	·034	·030	·026	·022	·018	·014	·010

* The choice of proper sign must be made as before.

TABLE III.

Correlation between Father's Stature and Son's Stature

Son's Stature $= x$.

Father's Stature $= y$.

	4'11"½—5'0"½	5'0"½—5'1"½	5'1"½—5'2"½	5'2"½—5'3"½	5'3"½—5'4"½	5'4"½—5'5"½	5'5"½—5'6"½	5'6"½—5'7"½	5'7"½—5'8"½	5'8"½—5'9"½	5'9"½—5'10"½	5'10"½—5'11"½	5'11"½—6'0"½	6'0"½—6'1"½	6'1"½—6'2"½	6'2"½—6'3"½	6'3"½—6'4"½	6'4"½—6'5"½	6'5"½—6'6"½	6'6"½—6'7"½	6'7"½—6'8"½	Totals
4'10"½—4'11"½	—	—	—	—	1	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·00
4'11"½—5'0"½	—	—	·25	·25	—	1	·5	1·5	—	—	—	—	—	—	—	—	—	—	—	—	—	3·50
5'0"½—5'1"½	—	—	·25	·25	1·5	·5	1	1·5	1	1	—	—	—	—	—	—	—	—	—	—	—	7·00
5'1"½—5'2"½	—	—	·5	2·75	3·75	2	2·25	5·25	2·5	—	—	—	—	—	—	—	—	—	—	—	—	19·00
5'2"½—5'3"½	·5	·5	·5	2·25	4	2·75	6·25	4·5	7·75	5	1	—	—	—	1	—	—	—	—	—	—	36·00
5'3"½—5'4"½	·5	—	2	2	4·25	7·5	11·5	12·75	11	5·5	3	3·25	·5	·25	—	—	—	—	—	—	—	64·00
5'4"½—5'5"½	1	·25	·25	3·75	7·75	12·5	11	20·75	10	11·75	6·25	5	3	·75	1·5	—	—	—	—	—	—	95·50
5'5"½—5'6"½	—	·25	·25	4·75	8·5	10·75	17·5	27·25	25	19·75	18·75	7·75	1	·5	1·5	—	—	—	—	—	—	143·50
5'6"½—5'7"½	—	1	·5	2·75	3·5	10·5	16·75	26·5	32·75	17·25	14·75	11·25	8	2·5	—	—	—	—	—	—	—	148·00
5'7"½—5'8"½	—	—	·5	1·25	2·25	5·5	15·5	20·5	25	25·75	20	17·25	9·75	6·5	6·25	2	1·25	·25	—	—	—	159·50
5'8"½—5'9"½	—	—	—	—	2·5	2·5	5·25	14	28·75	29	21	16·5	11·25	7	2·75	2	·25	·25	—	—	—	143·00
5'9"½—5'10"½	—	—	—	·25	·75	2·5	2	15·25	12·5	21·75	20	20·25	12·75	6·5	2·5	—	—	—	1	·5	·5	119·00
5'10"½—5'11"½	—	—	—	·25	1·25	—	2·5	3·25	9·5	10·5	13·5	11·25	10·5	8	6·5	2·5	·5	—	—	—	—	80·00
5'11"½—6'0"½	—	—	—	—	—	—	1	1	9	4·5	6·25	9	8·5	7·25	3·25	·75	1·5	·5	·25	·25	—	53·00
6'0"½—6'1"½	—	—	—	—	—	—	—	1	2·25	2·25	3·5	5	3·75	3·25	2·75	2·25	1	·5	·75	·25	—	28·50
6'1"½—6'2"½	—	—	—	—	—	—	—	—	—	—	1·5	1	·5	·5	—	·5	—	—	—	—	—	4·00
6'2"½—6'3"½	—	—	—	—	—	—	—	—	—	1	1	1	—	2·5	—	—	—	—	—	—	—	5·50
Totals	2·00	2·00	5·00	20·50	41·00	60·00	93·00	155·00	177·00	155·00	130·50	108·50	69·50	45·50	28·00	10·00	4·50	1·50	2·00	1·00	·50	1112

$m_x = 5'8''\cdot 65 \pm \cdot 06$ $\sigma_x = 2\cdot 74 \pm \cdot 04$ $r_{xy} = \cdot 515 \pm \cdot 015$

$m_y = 5'7''\cdot 70 \pm \cdot 05$ $\sigma_y = 2\cdot 72 \pm \cdot 04$

(N.B. No correction for grouping applied.)

TABLE IV.

Correlation between Cube of Father's Stature and Cube of Son's Stature.

(Son's Stature)$^3 \equiv u$. *Unit* 1000 *cubic inches.*

(Father's Stature) $\equiv w$. *Unit* 1000 *cubic inches.*

	210—220	220—230	230—240	240—250	250—260	260—270	270—280	280—290	290—300	300—310	310—320	320—330	330—340	340—350	350—360	360—370	370—380	380—390	390—400	400—410	410—420	420—430	430—440	440—450	450—460	460—470	470—480	480—490	490—500	500—510	Totals
200—210	—	—	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
210—220	—	—	—	—	—	—	1	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
220—230	—	—	—	1	—	1	—	2	—	2	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	8
230—240	—	—	—	1	1	2	2	1	1	3	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	13
240—250	—	—	—	—	2	3	—	—	3	4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	12
250—260	1	—	1	—	5	3	1	9	1	3	6	5	3	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	39
260—270	—	—	—	2	2	5	6	9	6	15	6	9	—	2	5	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	69
270—280	1	—	—	—	3	6	7	4	8	12	2	11	1	3	3	2	2	—	1	1	—	—	—	—	—	—	—	—	—	—	67
280—290	—	—	1	1	5	8	6	11	7	22	12	17	9	13	9	1	—	1	1	1	—	—	—	—	—	—	—	—	—	—	125
290—300	—	—	—	1	3	3	5	9	12	13	9	7	4	5	5	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	77
300—310	—	1	1	—	2	3	7	8	10	24	22	15	6	4	12	8	2	4	—	—	—	—	—	—	—	—	—	—	—	—	129
310—320	—	—	—	—	1	2	2	6	10	11	9	18	10	8	8	8	5	1	3	3	—	2	1	—	—	—	—	—	—	—	108
320—330	—	—	—	—	—	2	1	4	10	12	19	26	12	7	16	6	4	4	5	2	—	2	—	1	—	—	—	—	—	—	133
330—340	—	—	—	—	—	1	3	1	4	4	5	16	7	7	14	2	6	7	1	—	—	—	—	—	—	—	1	—	—	—	79
340—350	—	—	—	—	—	—	—	—	3	10	3	14	4	2	9	6	2	1	2	2	—	—	—	—	—	—	1	—	—	1	60
350—360	—	—	—	—	1	1	—	1	3	3	8	10	10	9	13	8	7	4	3	4	—	1	1	—	—	—	—	—	—	—	87
360—370	—	—	—	—	—	—	—	—	—	1	4	5	2	3	5	1	3	3	4	3	1	—	—	—	—	—	—	—	—	—	35
370—380	—	—	—	—	—	—	—	1	—	1	4	1	—	1	4	3	3	4	1	—	—	—	1	1	—	—	—	—	—	—	25
380—390	—	—	—	—	—	—	—	—	—	1	1	2	3	1	2	1	1	4	1	3	2	1	1	—	—	1	—	1	—	—	26
390—400	—	—	—	—	—	—	—	—	—	—	—	—	—	3	—	3	—	—	—	—	1	1	—	—	—	—	—	—	—	—	8
400—410	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
410—420	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	2
420—430	—	—	—	—	—	—	—	—	—	—	—	—	1	—	1	—	—	1	—	1	—	—	—	—	—	—	—	—	—	—	4
Totals	2	1	3	6	25	40	43	66	78	142	113	157	73	68	107	52	36	35	22	21	4	7	4	2	—	1	2	1	—	1	1112

$m_u = 325600 \pm 800$ $\sigma_u = 38970 \pm 560$ $r_{uw} = {\cdot}514 \pm {\cdot}015$

$m_w = 312000 \pm 700$ $\sigma_w = 36680 \pm 520$

(N.B. No correction for grouping applied.)

The correlations as found by the product-moment method were respectively:

$$r_{xy} = \cdot 515 \pm \cdot 015, \quad r_{uw} = \cdot 514 \pm \cdot 015.$$

Here $\lambda_1'' = \lambda_2'' = 2, \quad v_x = \cdot 03997, \quad v_y = \cdot 04014.$

Applying formula (13) we have

$$r_{uw} = \cdot 515 \left[1 - \tfrac{1}{4}(\cdot 04014)^2\, 4 - \tfrac{1}{4}(\cdot 03997)^2\, 4 + \tfrac{1}{2}(\cdot 04014)(\cdot 03997) \times \cdot 515\right]$$
$$= \cdot 514,$$

agreeing exactly with correlation table method. It is clear that in such a case nothing whatever is to be gained by using a function of x instead of x in correlating.

(3) *Frequency Distributions.* We now pass to another matter which has considerably troubled certain statisticians and to some extent the trouble may be justified. Namely it is argued that if x follow a normal distribution it is not possible for $u = \phi_1(x)$ to do so also. This is clearly correct theoretically, but we have not yet met with an attempt to estimate the degree of divergence from the exact values when, knowing that x follows a normal distribution, we proceed to calculate the frequencies of u from a normal distribution also. We shall now investigate this point.

First Method. From (4) and (6) we get[*]

$$\begin{aligned}\frac{u'}{\sigma_u} = \pm \Bigg[& \frac{x'}{\sigma_x} + \tfrac{1}{2} v_x \lambda_1'' \left\{\left(\frac{x'}{\sigma_x}\right)^2 - \sqrt{{}_x\beta_1}\,\frac{x'}{\sigma_x} - 1\right\} + \tfrac{1}{6} v_x^2 \lambda_1''' \left\{\left(\frac{x'}{\sigma_x}\right)^3 - {}_x\beta_2 \frac{x'}{\sigma_x} - \sqrt{{}_x\beta_1}\right\} \\ & - \tfrac{1}{8} v_x^2 \lambda_1''^2 \left\{2\sqrt{{}_x\beta_1}\left(\left(\frac{x'}{\sigma_x}\right)^2 - 1\right) + ({}_x\beta_2 - 3\,{}_x\beta_1 - 1)\frac{x'}{\sigma_x}\right\} \\ & - \tfrac{1}{16} v_x^3 \lambda_1''^3 \left\{\left(\frac{x'}{\sigma_x}\right)^2 ({}_x\beta_2 - 3\,{}_x\beta_1 - 1) - \frac{x'}{\sigma_x}\sqrt{{}_x\beta_1}\,(3\,{}_x\beta_2 - 5\,{}_x\beta_1 - 3) - {}_x\beta_2 + {}_x\beta_1 + 1\right\} \\ & - \tfrac{1}{12} v_x^3 \lambda_1'' \lambda_1''' \left\{\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^3 + {}_x\beta_2\left(\frac{x'}{\sigma_x}\right)^2 + ({}_x\beta_3' - 3\,{}_x\beta_2\sqrt{{}_x\beta_1} - \sqrt{{}_x\beta_1})\frac{x'}{\sigma_x} - {}_x\beta_2 - {}_x\beta_1\right\} \\ & - \tfrac{1}{24} v_x^3 \lambda_1^{\text{iv}} \left\{{}_x\beta_2 + {}_x\beta_3'\,\frac{x'}{\sigma_x} - \left(\frac{x'}{\sigma_x}\right)^4\right\} + \text{terms in } v_x^4 \text{ and higher powers}\Bigg] \quad \ldots(15).\end{aligned}$$

Differentiating (15)

$$\begin{aligned}\frac{u'}{\sigma_u} = \pm \frac{dx'}{\sigma_x} \Bigg[& 1 + \tfrac{1}{2} v_x \lambda_1'' \left\{2\,\frac{x'}{\sigma_x} - \sqrt{{}_x\beta_1}\right\} + \tfrac{1}{6} v_x^2 \lambda_1''' \left\{3\left(\frac{x'}{\sigma_x}\right)^2 - {}_x\beta_2\right\} \\ & - \tfrac{1}{8} v_x^2 \lambda_1''^2 \left\{4\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right) + {}_x\beta_2 - 3\,{}_x\beta_1 - 1\right\} \\ & - \tfrac{1}{16} v_x^3 \lambda_1''^3 \left\{2({}_x\beta_2 - 3\,{}_x\beta_1 - 1)\frac{x'}{\sigma_x} - \sqrt{{}_x\beta_1}\,(3\,{}_x\beta_2 - 5\,{}_x\beta_1 - 3)\right\} \\ & - \tfrac{1}{12} v_x^3 \lambda_1'' \lambda_1''' \left\{3\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^2 + 2\,{}_x\beta_2\,\frac{x'}{\sigma_x} + {}_x\beta_3' - 3\,{}_x\beta_2\sqrt{{}_x\beta_1} - \sqrt{{}_x\beta_1}\right\} \\ & + \tfrac{1}{24} v_x^3 \lambda_1^{\text{iv}} \left\{4\left(\frac{x'}{\sigma_x}\right)^3 - {}_x\beta_3'\right\} + \text{terms in } v_x^4 \text{ and higher powers}\Bigg] \ldots(16).\end{aligned}$$

---

[*] See footnote on p. 214.

Squaring (15)

$$\left(\frac{u'}{\sigma_u}\right)^2 = \left(\frac{x'}{\sigma_x}\right)^2 + v_x\lambda_1''\left\{\left(\frac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^2 - \frac{x'}{\sigma_x}\right\}$$

$$+ \tfrac{1}{3}v_x^2\lambda_1'''\left\{\left(\frac{x'}{\sigma_x}\right)^4 - {}_x\beta_2\left(\frac{x'}{\sigma_x}\right)^2 - \sqrt{{}_x\beta_1}\frac{x'}{\sigma_x}\right\}$$

$$+ \tfrac{1}{4}v_x^2\lambda_1''^2\left\{\left(\frac{x'}{\sigma_x}\right)^4 - 4\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^3 - ({}_x\beta_2 - 4{}_x\beta_1 + 1)\left(\frac{x'}{\sigma_x}\right)^2 + 4\sqrt{{}_x\beta_1}\frac{x'}{\sigma_x} + 1\right\}$$

$$- \tfrac{1}{8}v_x^3\lambda_1''^3\left\{2\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^4 + 2({}_x\beta_2 - 4{}_x\beta_1 - 1)\left(\frac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1}(4{}_x\beta_2 - 8{}_x\beta_1)\left(\frac{x'}{\sigma_x}\right)^2\right.$$

$$\left. - 2({}_x\beta_2 - 4{}_x\beta_1 - 1)\frac{x'}{\sigma_x} + 2\sqrt{{}_x\beta_1}\right\}$$

$$+ \tfrac{1}{6}v_x^3\lambda_1''\lambda_1'''\left\{\left(\frac{x'}{\sigma_x}\right)^5 - 2\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^4 - (2{}_x\beta_2 + 1)\left(\frac{x'}{\sigma_x}\right)^3\right.$$

$$\left. - ({}_x\beta_3' - 4{}_x\beta_2\sqrt{{}_x\beta_1})\left(\frac{x'}{\sigma_x}\right)^2 + 2({}_x\beta_1 + {}_x\beta_2)\frac{x'}{\sigma_x} + \sqrt{{}_x\beta_1}\right\}$$

$$- \tfrac{1}{12}v_x^3\lambda_1^{\text{IV}}\left\{{}_x\beta_2\frac{x'}{\sigma_x} + {}_x\beta_3'\left(\frac{x'}{\sigma_x}\right)^2 - \left(\frac{x'}{\sigma_x}\right)^5\right\}$$

+ terms in v_x^4 and higher powers ....(17).

Hence raising $(u'/\sigma_u)^2$ to the exponential and expanding the right-hand side we find:

$$e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2} = e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2}$$

$$1 - \tfrac{1}{2}v_x\lambda_1''\left\{\left(\frac{x'}{\sigma_x}\right)^3 - \sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^2 - \frac{x'}{\sigma_x}\right\}$$

$$- \tfrac{1}{6}v_x^2\lambda_1'''\left\{\left(\frac{x'}{\sigma_x}\right)^4 - {}_x\beta_2\left(\frac{x'}{\sigma_x}\right)^2 - \sqrt{{}_x\beta_1}\frac{x'}{\sigma_x}\right\}$$

$$+ \tfrac{1}{8}v_x^2\lambda_1''^2\left\{\left(\frac{x'}{\sigma_x}\right)^6 - 2\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^5 + ({}_x\beta_1 - 3)\left(\frac{x'}{\sigma_x}\right)^4 + 6\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^3\right.$$

$$\left. + ({}_x\beta_2 - 4{}_x\beta_1 + 2)\left(\frac{x'}{\sigma_x}\right)^2 - 4\sqrt{{}_x\beta_1}\frac{x'}{\sigma_x} - 1\right\}$$

$$- \tfrac{1}{48}v_x^3\lambda_1''^3\left\{\left(\frac{x'}{\sigma_x}\right)^9 - 3\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^8 + (3{}_x\beta_1 - 6)\left(\frac{x'}{\sigma_x}\right)^7 + \sqrt{{}_x\beta_1}(21 - {}_x\beta_1)\left(\frac{x'}{\sigma_x}\right)^6\right.$$

$$+ (3{}_x\beta_2 - 27{}_x\beta_1 + 9)\left(\frac{x'}{\sigma_x}\right)^5 - 3\sqrt{{}_x\beta_1}({}_x\beta_2 - 4{}_x\beta_1 + 12)\left(\frac{x'}{\sigma_x}\right)^4$$

$$- (9{}_x\beta_2 - 48{}_x\beta_1 + 1)\left(\frac{x'}{\sigma_x}\right)^3 + \sqrt{{}_x\beta_1}(12{}_x\beta_2 - 24{}_x\beta_1 + 15)\left(\frac{x'}{\sigma_x}\right)^2$$

$$\left. + 3(2{}_x\beta_2 - 8{}_x\beta_1 - 1)\frac{x'}{\sigma_x} - 6\sqrt{{}_x\beta_1}\right\}$$

$$+\tfrac{1}{12}v_x^3\lambda_1''\lambda_1'''\Bigl\{\left(\frac{x'}{\sigma_x}\right)^7-\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^6-({}_x\beta_2+2)\left(\frac{x'}{\sigma_x}\right)^5+\sqrt{{}_x\beta_1}\,({}_x\beta_2+{}_x\beta_3+1)\left(\frac{x'}{\sigma_x}\right)^4$$

$$+(3{}_x\beta_2+{}_x\beta_1+1)\left(\frac{x'}{\sigma_x}\right)^3+({}_x\beta_3'-4{}_x\beta_2\sqrt{{}_x\beta_1}+\sqrt{{}_x\beta_1})\left(\frac{x'}{\sigma_x}\right)^2$$

$$-2\,({}_x\beta_1+{}_x\beta_2)\frac{x'}{\sigma_x}-\sqrt{{}_x\beta_1}\Bigr\}$$

$$-\tfrac{1}{24}v_x^3\lambda_1^{\text{iv}}\Bigl\{\left(\frac{x'}{\sigma_x}\right)^5-{}_x\beta_3'\left(\frac{x'}{\sigma_x}\right)^2-{}_x\beta_2\frac{x'}{\sigma_x}$$

$$+\text{ terms in } v_x^4 \text{ and higher powers}\Bigr] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(18).$$

Multiplying (16) and (18) we get

$$e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u}=\pm\, e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2}\frac{dx'}{\sigma_x}$$

$$\Bigl[1-\tfrac{1}{2}v_x\lambda_1''\Bigl\{\left(\frac{x'}{\sigma_x}\right)^3-\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^2-3\left(\frac{x'}{\sigma_x}\right)+\sqrt{{}_x\beta_1}\Bigr\}$$

$$+\tfrac{1}{8}v_x^2\lambda_1''^2\Bigl\{\left(\frac{x'}{\sigma_x}\right)^6-2\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^5+({}_x\beta_1-7)\left(\frac{x'}{\sigma_x}\right)^4+12\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^3$$

$$+({}_x\beta_2-6{}_x\beta_1+6)\left(\frac{x'}{\sigma_x}\right)^2-10\sqrt{{}_x\beta_1}\frac{x'}{\sigma_x}-{}_x\beta_2+3{}_x\beta_1\Bigr\}$$

$$-\tfrac{1}{6}v_x^2\lambda_1'''\Bigl\{\left(\frac{x'}{\sigma_x}\right)^4-({}_x\beta_2+3)\left(\frac{x'}{\sigma_x}\right)^2-\sqrt{{}_x\beta_1}\frac{x'}{\sigma_x}+{}_x\beta_2\Bigr\}$$

$$-\tfrac{1}{48}v_x^3\lambda_1''^3\Bigl\{\left(\frac{x'}{\sigma_x}\right)^9-3\sqrt{{}_x\beta_1}\left(\frac{x}{\sigma_x}\right)^8+3\,({}_x\beta_1-4)\left(\frac{x'}{\sigma_x}\right)^7+\sqrt{{}_x\beta_1}\,(36-{}_x\beta_1)\left(\frac{x'}{\sigma_x}\right)^6$$

$$+3\,({}_x\beta_2-13{}_x\beta_1+9)\left(\frac{x'}{\sigma_x}\right)^5-3\sqrt{{}_x\beta_1}\,({}_x\beta_2-5{}_x\beta_1+31)\left(\frac{x'}{\sigma_x}\right)^4$$

$$-(18{}_x\beta_2-111{}_x\beta_1+10)\left(\frac{x'}{\sigma_x}\right)^3+\sqrt{{}_x\beta_1}\,(18{}_x\beta_2-45{}_x\beta_1+54)\left(\frac{x'}{\sigma_x}\right)^2$$

$$+3\,(5{}_x\beta_2-21{}_x\beta_1-2)\frac{x'}{\sigma_x}-3\sqrt{{}_x\beta_1}\,(3{}_x\beta_2-5{}_x\beta_1)\Bigr\}$$

$$+\tfrac{1}{12}v_x^3\lambda_1''\lambda_1'''\Bigl\{\left(\frac{x'}{\sigma_x}\right)^7-\sqrt{{}_x\beta_1}\left(\frac{x'}{\sigma_x}\right)^6-({}_x\beta_2+7)\left(\frac{x'}{\sigma_x}\right)^5+\sqrt{{}_x\beta_1}\,({}_x\beta_2+5)\left(\frac{x'}{\sigma_x}\right)^4$$

$$+(6{}_x\beta_2+{}_x\beta_1+4)\left(\frac{x'}{\sigma_x}\right)^3+({}_x\beta_3'-6{}_x\beta_2\sqrt{{}_x\beta_1})\left(\frac{x'}{\sigma_x}\right)^2-(5{}_x\beta_2+3{}_x\beta_1)\frac{x'}{\sigma_x}$$

$$-{}_x\beta_3'+3{}_x\beta_2\sqrt{{}_x\beta_1}\Bigr\}$$

$$-\tfrac{1}{24}v_x^3\lambda_1^{\text{iv}}\Bigl\{\left(\frac{x'}{\sigma_x}\right)^5-4\left(\frac{x'}{\sigma_x}\right)^3-{}_x\beta_3'\left(\frac{x'}{\sigma_x}\right)^2-{}_x\beta_2\frac{x'}{\sigma_x}+{}_x\beta_3'\Bigr\}$$

$$+\text{ terms in } v_x^4 \text{ and higher powers}\Bigr] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(19).$$

For the particular case of x having normal frequency

$$e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u} = \pm\, e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2}\frac{dx'}{\sigma_x}$$

$$\begin{aligned}\Bigg[1 &- \tfrac{1}{2} v_x \lambda_1'' \left\{\left(\frac{x'}{\sigma_x}\right)^3 - 3\left(\frac{x'}{\sigma_x}\right)\right\} \\ &+ \tfrac{1}{8} v_x^2 \lambda_1''^2 \left\{\left(\frac{x'}{\sigma_x}\right)^6 - 7\left(\frac{x'}{\sigma_x}\right)^4 + 9\left(\frac{x'}{\sigma_x}\right)^2 - 3\right\} \\ &- \tfrac{1}{6} v_x^2 \lambda_1''' \left\{\left(\frac{x'}{\sigma_x}\right)^4 - 6\left(\frac{x'}{\sigma_x}\right)^2 + 3\right\} \\ &- \tfrac{1}{48} v_x^3 \lambda_1''^3 \left\{\left(\frac{x'}{\sigma_x}\right)^9 - 12\left(\frac{x'}{\sigma_x}\right)^7 + 36\left(\frac{x'}{\sigma_x}\right)^5 - 64\left(\frac{x'}{\sigma_x}\right)^3 + 39\left(\frac{x'}{\sigma_x}\right)\right\} \\ &+ \tfrac{1}{12} v_x^3 \lambda_1'' \lambda_1''' \left\{\left(\frac{x'}{\sigma_x}\right)^7 - 10\left(\frac{x'}{\sigma_x}\right)^5 + 22\left(\frac{x'}{\sigma_x}\right)^3 - 15\left(\frac{x'}{\sigma_x}\right)\right\} \\ &- \tfrac{1}{24} v_x^3 \lambda_1^{\text{iv}} \left\{\left(\frac{x'}{\sigma_x}\right)^5 - 4\left(\frac{x'}{\sigma_x}\right)^3 - 3\left(\frac{x'}{\sigma_x}\right)\right\} \\ &+ \text{terms in } v_x^4 \text{ and higher powers}\Bigg] \quad \dots\dots\dots\dots\dots\dots (20).\end{aligned}$$

Multiplying $\frac{1}{\sqrt{2\pi}}$ into (20) and integrating this product between corresponding values of u' and x', i.e. u_1' and u_2' corresponding to x_1' and x_2', we obtain (21).

All the integrations on the right can be found by tables of the incomplete normal moment functions using m_n in the significance of the book of *Tables**, supposing x_1 to be zero and m_n to stand for $m_n(x_2)$.

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int_{u_1'}^{u_2'} & e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u} \\ = \pm\,\{m_0 &- \tfrac{1}{2} v_x \lambda_1'' (2m_2 - 3m_1) \\ &+ \tfrac{1}{8} v_x^2 \lambda_1''^2 (15m_6 - 21m_4 + 9m_2 - 3m_0) \\ &- \tfrac{1}{6} v_x^2 \lambda_1''' (3m_4 - 6m_2 + 3m_0) \\ &- \tfrac{1}{48} v_x^3 \lambda_1''^3 (384m_9 - 576m_7 + 288m_5 - 128m_3 + 39m_1) \\ &+ \tfrac{1}{12} v_x^3 \lambda_1'' \lambda_1''' (48m_7 - 80m_5 + 44m_3 - 15m_1) \\ &- \tfrac{1}{24} v_x^3 \lambda_1^{\text{iv}} (8m_5 - 8m_3 - 3m_1) \\ &+ \text{terms in } v_x^4 \text{ and higher powers}\} \quad \dots\dots\dots\dots\dots\dots (21).\end{aligned}$$

(21) gives the frequency of normal distribution with regard to u, between u_1' and u_2', having its mode at the mean of u, i.e. $u' = 0$†. True frequency in that interval being m_0, the total value of the remaining terms represents an excess beyond normality. This excess divided by m_0 and multiplied by 100 will give the

---

* See *Tables for Statisticians and Biometricians*, Part I. pp. 22—28, Table IX.

† This is the usual manner of fitting a normal curve to a series of observations, i.e. we make the mean of observations and origin of normal curve coincide, and use the second moment coefficient of the observations (σ_u^2) about that mean.

percentage of deviation from normality or as we may say percentage errors. Table V gives these errors for the special case of $u = kx^3$ and $v_x = ·04$.

In Table V, the first two columns give the corresponding values of x'/σ_x and u'/σ_u at the integral limits. The third column gives the exact frequency m_0. The

TABLE V.

Comparison of Exact Frequency and that of Normal Curve having its Mode at the Mean of u. Calculated by (21) for $u = kx^3$ and $v_x = 0·4$.

$\frac{x'}{\sigma_x}$	$\frac{u'}{\sigma_u}$	m_0	v_x	v_x^2	v_x^3	Combined Results
$\frac{x_1'}{\sigma_x} - \frac{x_2'}{\sigma_x}$	$\frac{u_1'}{\sigma_u} - \frac{u_2'}{\sigma_u}$					
0— ·1	− ·03987253 ·06020848	= ·0398278	+ ·00023836	− ·00015794	− ·00000145	= ·0399068
0— ·5	− ·03987253 ·4685687	= ·1914625	+ ·00539574	− ·00061905	− ·00002843	= ·1962108
0— 1·0	− ·03987253 ·9973448	= ·3413447	+ ·01595770	− ·00064526	− ·00005532	= ·3566018
0— 1·5	− ·03987253 1·5468547	= ·4331928	+ ·02243357	− ·00063140	− ·00004353	= ·4549514
0— 2·5	− ·03987253 2·7096705	= ·4937903	+ ·01963864	− ·00096040	+ ·00007187	= ·5125404
0—− ·1	− ·03987253 − ·1391557	= ·0398278	− ·00023836	− ·00015794	+ ·00000145	= ·0394330
0—− ·5	− ·03987253 − ·5283774	= ·1914625	− ·00539574	− ·00061905	+ ·00002843	= ·1854761
0—− 1·0	− ·03987253 − ·9973448	= ·3413447	− ·01595770	− ·00064526	+ ·00005532	= ·3247971
0—− 1·5	− ·03987253 − 1·4471734	= ·4331928	− ·02243357	− ·00063140	+ ·00004353	= ·4101714
0—− 2·5	− ·03987253 − 2·2910090	= ·4937903	− ·01963864	− ·00096040	− ·00007187	= ·4731194

fourth, fifth and sixth columns give the contributions of the terms in v_x, v_x^2 and v_x^3 to the total value of the integral

$$\frac{1}{\sqrt{2\pi}}\int_{u_1'}^{u_2'} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u}$$

as set out in (21).

The last column gives the combined results; and the comparison is to be made between the m_0 column and the combined result in the last column. A comparison of these two columns indicates that for many practical purposes a knowledge of the v_x term would be adequate for correction of the frequency. But to this degree of approximation:

$$\frac{1}{\sqrt{2\pi}}\int_{u_1'}^{u_2'} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u} = \pm\frac{1}{\sqrt{2\pi}}\int_{x_1'}^{x_2'} e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2} dx'$$

$$\pm \tfrac{1}{2} v_x\lambda_1''\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2}\left(\left(\frac{x'}{\sigma_x}\right)^2-1\right)\right]_{x_1'}^{x_2'}.$$

The first integral on the right is the probability integral equal to (in the notation of the *Tables for Statisticians*) $\frac{1}{2}(1+\alpha_{x_2'})-\frac{1}{2}(1+\alpha_{x_1'})$. The term in square brackets is $\sqrt{\frac{3}{2}}\,\tau_3(x'/\sigma_x)$, where τ_3 is the tetrachoric function of the third order and will be found tabled in the book of *Tables for Statisticians*, Table XXIX, pp. 42—51. Thus we have

$$\frac{1}{\sqrt{2\pi}}\int_{u_1'}^{u_2'} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u} = \tfrac{1}{2}(1+\alpha_{x_2'})-\tfrac{1}{2}(1+\alpha_{x_1'}) \pm \sqrt{\tfrac{3}{2}}\, v_x\lambda_1''\left(\tau_3\left(\frac{x_2'}{\sigma_x}\right)-\tau_3\left(\frac{x_1'}{\sigma_x}\right)\right)$$

$$= \tfrac{1}{2}(1-\alpha_{x_1'})-\tfrac{1}{2}(1-\alpha_{x_2'}) \pm \sqrt{\tfrac{3}{2}}\, v_x\lambda_1''\left(\tau_3\left(\frac{x_2'}{\sigma_x}\right)-\tau_3\left(\frac{x_1'}{\sigma_x}\right)\right) \quad \dots (22),$$

and the approximation can thus be found by use of the tables of the probability integral and that of the third tetrachoric function. Some but not so great a gain can be obtained by expressing the coefficients of v_x^2 and v_x^3 in terms of the tetrachoric functions instead of the incomplete moment functions.

It is easy to obtain from τ_3 an appreciation of the maximum error due to this correction.

If we take any two successive values in the first column of the Table XXIX for $\frac{1}{2}(1-\alpha_{x_1'})$, $\frac{1}{2}(1-\alpha_{x_2'})$ then $\tau_3\left(\frac{x_1'}{\sigma_x}\right)$, $\tau_3\left(\frac{x_2'}{\sigma_x}\right)$ will be given in the corresponding row in the τ_3 column. Since the difference of the former two is constant the region of maximum difference in the τ_3 column will give maximum error.

So long as the central region of the Table covers the difference, ·00082 will be found to be the maximum. The corresponding values of $\frac{1}{2}(1-\alpha)$ will run from ·174 to ·146 with mean of ·160 for which $\frac{x'}{\sigma_x} = 1{\cdot}08$.

Hence maximum error will be roughly

$$100\sqrt{\tfrac{3}{2}}\,.\,v_x\lambda_1'' \times {\cdot}00082/{\cdot}001 \equiv E_{\text{max.}} \text{ say.}$$

Let us take again $v_x = {\cdot}04$, then we find:

A. $u = kx^3$, $\lambda_1'' = 2$: $|E_{\text{max.}}| = 8{\cdot}0\,\%$.

B. $u = k\sqrt{x}$, $\lambda_1'' = -\frac{1}{2}$: $|E_{\text{max.}}| = 2{\cdot}0\,\%$.

C. $u = \frac{k}{x}$, $\lambda_1'' = -2$: $|E_{\text{max.}}| = 8{\cdot}0\,\%$.

D. $u = k\log x$, $\lambda_1'' = -1$: $|E_{\text{max.}}| = 4{\cdot}0\,\%$.

If we choose $\frac{x_1'}{\sigma_x} = 0$, so that we have:

$$\tfrac{1}{2}(1 - \alpha_{x_1'}) = {\cdot}5000, \quad \tau_3\left(\frac{x_1'}{\sigma_x}\right) = -{\cdot}16287,$$

we must find the value of x_2'/σ_x which gives the maximum error; this means that we must look for the maximum value of

$$\left\{\tau\left(\frac{x_2'}{\sigma_x}\right) + {\cdot}16287\right\}\Big/\left\{{\cdot}500 - \tfrac{1}{2}(1 - \alpha_{x_2'})\right\}.$$

We easily deduce that

$$\frac{{\cdot}06304 + {\cdot}16287}{{\cdot}500 - {\cdot}073} = 52{\cdot}906,$$

corresponding to $x_2'/\sigma_x = 1{\cdot}45$, gives us the maximum.

Hence $\qquad$ max. error $\fallingdotseq 100\sqrt{\tfrac{3}{2}}\,v_x\lambda_1'' \times 52{\cdot}906$.

Let $v_x = {\cdot}04$. Then:

A. $u = kx^3$, $\lambda_1'' = 2$: max. error $\fallingdotseq 5{\cdot}18\,\%$.

B. $u = k\sqrt{x}$, $\lambda_1'' = -\tfrac{1}{2}$: max. error $\fallingdotseq 1{\cdot}30\,\%$.

C. $u = \frac{k}{x}$, $\lambda_1'' = -2$: max. error $\fallingdotseq 5{\cdot}18\,\%$.

D. $u = k\log x$, $\lambda_1'' = -1$: max. error $\fallingdotseq 2{\cdot}59\,\%$.

Further we can predict that in the case of different functions the percentage errors will be approximately proportional to their λ_1'''s.

Difference of τ_3 vanishes at the point

$$\tfrac{1}{2}(1 - \alpha) = {\cdot}0415, \quad \text{or} \quad x'/\sigma_x \fallingdotseq 1{\cdot}73.$$

At that point the error will be very small. In the case of either "tail," the further we go, the greater the percentage error will be; thus there is no maximum in the case of "tail" evaluation. This is, however, of small importance as the absolute frequency in the tails becomes for practical purposes insensible.

Another use may be made of Equation (22). It is clear that the tetrachoric term gives the fundamental correction on the normal distribution of u'/σ_u. But the corrections for any two functions will have a ratio depending on the ratio of their λ_1'''s. Thus we see that for the same range of values of x_1'/σ_x to x_2'/σ_x the correction might be obtained approximately by multiplying those obtained for any standard function $({}_0v_x, {}_0\lambda_1'')$ by the ratio $v_x\lambda_1''/({}_0v_x, {}_0\lambda_1'')$, paying due regard to the signs of ${}_0\lambda_1''$ and λ_1''.

From (22) we deduce at once that

$$\text{True Frequency} = \frac{1}{\sqrt{2\pi}}\int_{u_1'}^{u_2'} e^{-\frac{1}{2}\frac{u'^2}{\sigma_u^2}}\frac{du'}{\sigma_u} \mp \sqrt{\tfrac{3}{2}}\,v_x\lambda_1''\left(\tau_3\left(\frac{x_2'}{\sigma_x}\right) - \tau_3\left(\frac{x_1'}{\sigma_x}\right)\right).$$

But the double sign is determined by that of $f_1'(1)$, being, now that it is transferred to the other side of the equation, the *opposite* sign to that of $f_1'(1)$. We may accordingly write the correction in the form

$$-\left\{\pm\sqrt{\tfrac{3}{2}}\,v_x\frac{f_1''(1)}{f_1'(1)}\right\}\left(\tau_3\left(\frac{x_2'}{\sigma_x}\right)-\tau_3\left(\frac{x_1'}{\sigma_x}\right)\right).$$

This shows that as far as the factor in curled brackets is concerned the sign of $f_1'(1)$ is indifferent. The factor therefore depends solely on the sign of $f_1''(1)$. If $f_1''(1)$ and the tetrachoric difference be of the same sign, the correction must be subtracted; if they be of opposite signs, it must be added. The sign of the tetrachoric difference depends entirely on the range of values selected for u_2' and u_1' and so for x_2' and x_1'; it can be positive, negative or zero. The sign of $f_1''(1)$ corresponds to an important physical relation. We have:

$$u_{\text{mean}}=f_1(1)+\tfrac{1}{2}v_x^2 f_1''(1),\ \text{nearly},$$
$$u_{\text{median}}=f_1(1),$$
$$u_{\text{mean}}-u_{\text{median}}=\tfrac{1}{2}v_x^2 f_1''(1).$$

Or, when $f_1''(1)$ is positive the mean is greater than the median and therefore greater than the mode (as in $u=kx^3$ and $u=k/x$); but if $f_1''(1)$ be negative (as in $u=k\sqrt{x}$ and $u=k\log x$) then the mode is greater than the mean.

Second Method. The same result can be reached in a different manner. On the left-hand side of the equation (21), if we only know the values of u_1' and u_2' corresponding respectively to x_1' and x_2' we can evaluate the above errors from the table of the probability integral simply.

The equation (7) gives the value of $\frac{u'}{\sigma_u}$ corresponding to $\frac{x'}{\sigma_x}$ but it is better calculated to slightly more accuracy for our present purpose.

Retaining in (4) and (6) a few more terms we find:

$$\frac{u'}{}=+\frac{\frac{x'}{\sigma_x}+\frac{1}{2}v_x\lambda_1''\left\{\left(\frac{x'}{\sigma_x}\right)^2-1\right\}+\frac{1}{6}v_x^2\lambda_1'''\left\{\left(\frac{x'}{\sigma_x}\right)^3-\sqrt{{}_x\beta_1}\right\}+\frac{1}{24}v_x^3\lambda_1^{\text{iv}}\left\{\left(\frac{x'}{\sigma_x}\right)^4-{}_x\beta_2\right\}+\frac{1}{120}v_x^4\lambda_1^{\text{v}}\left\{\left(\frac{x'}{\sigma_x}\right)^5-{}_x\beta_3'\right\}}{\left\{\begin{array}{l}1+v_x\lambda_1''\sqrt{{}_x\beta_1}+\frac{1}{3}v_x^2\lambda_1'''{}_x\beta_2+\frac{1}{4}v_x^2\lambda_1''^2({}_x\beta_2-1)+\frac{1}{12}v_x^3\lambda_1^{\text{iv}}{}_x\beta_3'+\frac{1}{6}v_x^3\lambda_1''\lambda_1'''\{{}_x\beta_3'-\sqrt{{}_x\beta_1}\}\\ \quad+\frac{1}{60}v_x^4\lambda_1^{\text{v}}{}_x\beta_4+\frac{1}{24}v_x^4\lambda_1''\lambda_1^{\text{iv}}({}_x\beta_4-{}_x\beta_2)+\frac{1}{36}v_x^4\lambda_1'''^2({}_x\beta_4-{}_x\beta_1)\end{array}\right\}^{\frac{1}{2}}}$$

In case of normal distribution of x' this reduces to:

$$\frac{u'}{\sigma_u}=\pm\frac{\frac{x'}{\sigma_x}+\frac{1}{2}v_x\lambda_1''\left\{\left(\frac{x'}{\sigma_x}\right)^2-1\right\}+\frac{1}{6}v_x^2\lambda_1'''\left(\frac{x'}{\sigma_x}\right)^3+\frac{1}{24}v_x^3\lambda_1^{\text{iv}}\left\{\left(\frac{x'}{\sigma_x}\right)^4-3\right\}+\frac{1}{120}v_x^4\lambda_1^{\text{v}}\left(\frac{x'}{\sigma_x}\right)^5}{\{1+v_x^2\lambda_1'''+\frac{1}{2}v_x^2\lambda_1''^2+\frac{1}{4}v_x^4\lambda_1^{\text{v}}+\frac{1}{2}v_x^4\lambda_1''\lambda_1^{\text{iv}}+\frac{5}{12}v_x^4\lambda_1'''^2\}^{\frac{1}{2}}}\ldots(23)\,bis.$$

From (23) *bis* we get Table VI.

Third Method. In the equation (23) corresponding values of $\frac{u'}{\sigma_u}$ and $\frac{x'}{\sigma_x}$ are independent of k and depend not on the magnitudes of x', u', $\bar{x}$, $\bar{u}$, σ_x, σ_u but on their relative values, i.e. $\frac{x'}{\sigma_x}$, $\frac{\bar{x}}{\sigma_x}$, $\frac{u'}{\sigma_u}$, $\frac{\bar{u}}{\sigma_u}$; therefore we can give k, $\bar{x}$, σ_x special values

TABLE VI.

Corresponding Values of $\frac{u'}{\sigma_u}$ *and* $\frac{x'}{\sigma_x}$.

$\frac{x'}{\sigma_x}$	$\frac{u'}{\sigma_u}$			
	$u=kx^3$	$u=k\sqrt{x}$	$u=\frac{k}{x}$	$u=k\log x$
+2·5	2·7096705	2·4487426	−2·2980793	2·3951220
+1·5	1·5468547	1·4871210	−1·4459359	1·4720481
+1·0	·9973448	·9995075	− ·9952935	·9973660
+ ·5	·4685687	·5071829	− ·5269799	·5134663
+ ·1	·0602085	·1098376	− ·1388953	·1194652
0	− ·0398725	·0100075	− ·0399337	·0199839
− ·1	− ·1391557	− ·0900224	− ·0598227	− ·0798962
− ·5	− ·5283774	− ·4921649	·4669918	− ·4834685
−1·0	− ·9973448	− ·9994880	·9950392	− ·9973022
−1·5	−1·4471734	−1·5121234	1·5455556	−1·5219539
−2·5	−2·2910090	−2·5540191	2·7199667	−2·6056007

without losing any generality, if we preserve the same value of v_x. Thus for $v_x = \cdot 04$:

A. $u = kx^3$.

Let $$k = 1,\ \bar{x} = 100,\ \sigma_x = 4.$$

Then by (2) $$\bar{u} = f_1(1) + \tfrac{1}{2} v_x^2 f_1''(1) = 1004800 \quad \text{.................}(24),$$

by (5)
$$\sigma_u = \cdot 04 \times 3000000 \sqrt{1 + \cdot 0016 \times 2 + \cdot 0016 \times 2 + \tfrac{5}{3} + \cdot 00000426667}$$
$$= 120383\cdot 643 \quad \text{..........................}(25),$$

$$120383\cdot 643 \frac{u'}{\sigma_u} + 1004800 = \left(4 \frac{x'}{\sigma_x} + 100\right)^3 \quad \text{..............}(26).$$

B. $u = k\sqrt{x}$.

Let $$k = 1,\ \bar{x} = 100,\ \sigma_x = 4.$$

Then $$\bar{u} = 9\cdot 997996987 \ . \ \text{..............................}(27),$$
$$\sigma_u = \cdot 2001405506 \ \text{..............................}(28),$$

$$\cdot 2001405506 \frac{u'}{\sigma_u} + 9\cdot 997996987 = \sqrt{4 \frac{x'}{\sigma_x} + 100} \quad \text{...........}(29).$$

C. $u = \frac{k}{x}$.

Let $$k = 1,\ \bar{x} = 100,\ \sigma_x = 4.$$

Then $$\bar{u} = \cdot 01001607741 \ \text{..........................}(30),$$
$$\sigma_u = \cdot 0004025869625 \ \text{.........................}(31),$$

$$\cdot 0004025869625 \frac{u'}{\sigma_u} + \cdot 01001607741 = \frac{1}{4 \frac{x'}{\sigma_x} + 100} \quad \text{...........}(32).$$

D. $u = k \log x$.

Let $$k = 1,\ \bar{x} = 100,\ \sigma_x = 4.$$

Then $$\bar{u} = 4\cdot604368260 \ldots\ldots\ldots\ldots(33),$$

$$\sigma_u = \cdot0401283337 \ldots\ldots\ldots\ldots(34),$$

$$\cdot0401283337 \frac{u'}{\sigma_u} + 4\cdot604368260 = \log_e\left(4\frac{x'}{\sigma_x} + 100\right) \quad \ldots\ldots\ldots(35).$$

By (26), (29), (32) and (35) we find on testing exactly the same results as those in Table VI. Moreover by these equations the inverse problem, i.e. to find the value of $\frac{x'}{\sigma_x}$ corresponding to the given value of $\frac{u'}{\sigma_u}$, can be solved quite easily.

I next proceeded to compute Tables for the above selected functions as typical cases, taking no longer total integrals from $x_1' = 0$ up to some value x_2', but integrals over various short ranges shown in Column (1) of the several sections of Table VII. While the errors for $\sqrt{x}$ do not reach anywhere 2 °/$_\circ$ and those for $\log x$ do not reach 4 °/$_\circ$, those for x^3 and $1/x$ may run up to 7 °/$_\circ$ or in one case 7·5 °/$_\circ$.

(4) *Effect of shifting the Origin of the Normal Curve for u.* An examination of Table VII indicates that while the normal curve for u gives a considerable number of percentage errors which are negligible from the statistical standpoint, there are others which we cannot overlook in practice. If we treat the v_x term as the most important, we have from (20)

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2} \frac{du'}{\sigma_u}\left\{1 \pm \tfrac{1}{2} v_x \lambda_1'' \left(\left(\frac{x'}{\sigma_x}\right)^3 - 3\left(\frac{x'}{\sigma_x}\right)\right)\right\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2} \frac{dx'}{\sigma_x} \ldots(36).$$

Hence by aid of (7), assuming normal distribution for x, we have to the same degree of approximation:

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2} \frac{du'}{\sigma_u}\left\{1 \pm \tfrac{1}{2} v_x \lambda_1'' \left(\left(\frac{u'}{\sigma_u}\right)^3 - 3\left(\frac{u'}{\sigma_u}\right)\right)\right\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2} \frac{dx'}{\sigma_x} \quad \ldots(37),$$

which is the true frequency.

Now it is not feasible to take a function $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u'+h}{\sigma_u+\epsilon}\right)^2} \frac{du'}{\sigma_u + \epsilon}$ and expanding in powers and products of h and ϵ to get by choice of h and ϵ a form like the above. The expansion consists of the term $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2} \frac{du'}{\sigma_u}$ multiplied by a polynomial in u'/σ_u and we have only two available constants h and ϵ to make the coefficients of $\frac{u'}{\sigma_u}$, $\left(\frac{u'}{\sigma_u}\right)^2$ and $\left(\frac{u'}{\sigma_u}\right)^3$ agree with the above values. We can get rid of $(u'/\sigma_u)^2$, but not of both the linear and the cubic terms. There is no point in getting rid of one of them, as they are approximately of the same order when u' and σ_u are not too different. Accordingly it is clear that we cannot find a normal curve which even to the term in v_x will give the true frequency of u'. One point,

TABLE VII.

Percentage Frequency Errors, when we use a Normal Curve having its Mean at the Mean of u.

A. $u = kx^3$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\,\Delta}{N}$
2·5 — 1·5	·0575817	·0605975	− ·0030158	− 4·9768
1·5 — 1·0	·0983494	·0918481	+ ·0065013	+ 7·0783
1·0 — ·5	·1603904	·1498822	+ ·0105082	+ 7·0110
·5 — ·1	·1563058	·1516347	+ ·0046711	+ 3·0805
·1 — 0	·0399079	·0398278	+ ·0000801	+ ·2011
0 — − ·1	·0394337	·0398278	− ·0003941	− ·9895
− ·1 — − ·5	·1460449	·1516347	− ·0075898	− 3·6864
− ·5 — − 1·0	·1393205	·1498822	− ·0105617	− 7·0467
− 1·0 — − 1·5	·0853740	·0918481	− ·0064741	− 7·0487
− 1·5 — − 2·5	·0629427	·0605975	+ ·0023452	+ 3·8701

B. $u = k\sqrt{x}$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\,\Delta}{N}$
2·5 — 1·5	·0613236	·0605975	+ ·0007261	+ 1·1982
1·5 — 1·0	·0902831	·0918481	− ·0015650	− 1·7039
1·0 — ·5	·1472387	·1498822	− ·0026435	− 1·7637
·5 — ·1	·1502559	·1516347	− ·0013788	− ·9093
·1 — 0	·0397385	·0398278	− ·0000893	− ·2242
0 — − ·1	·0398577	·0398278	+ ·0000299	+ ·0751
− ·1 — − ·5	·1528334	·1516347	+ ·0011987	+ ·7905
− ·5 — − 1·0	·1525221	·1498822	+ ·0026399	+ 1·7613
− 1·0 — − 1·5	·0935280	·0918481	+ ·0016799	+ 1·8290
− 1·5 — − 2·5	·0599269	·0605975	− ·0006706	− 1·1066

C. $u = \frac{k}{x}$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\,\Delta}{N}$
2·5 — 1·5	·0633189	·0605975	+ ·0027214	+ 4·4909
1·5 — 1·0	·0856997	·0918481	− ·0061484	− 6·6941
1·0 — ·5	·1393065	·1498822	− ·0105757	− 7·0560
·5 — ·1	·1456626	·1516347	− ·0059721	− 3·9385
·1 — 0	·0393065	·0398278	− ·0005213	− 1·3089
0 — − ·1	·0397787	·0398278	− ·0000491	− ·1233
− ·1 — − ·5	·1558956	·1516347	+ ·0042609	+ 2·8100
− ·5 — − 1·0	·1603942	·1498822	+ ·0105120	+ 7·0135
− 1·0 — − 1·5	·0987526	·0918481	+ ·0069045	+ 7·5173
− 1·5 — − 2·5	·0578416	·0605975	− ·0027559	− 4·5479

D. $u = k \log x$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\,\Delta}{N}$
2·5 — 1·5	·0621966	·0605975	+ ·0015991	+ 2·6389
1·5 — 1·0	·0887895	·0918481	− ·0030586	− 3·3301
1·0 — ·5	·1445191	·1498822	− ·0053631	− 3·5782
·5 — ·1	·1486408	·1516347	− ·0029939	− 1·9744
·1 — 0	·0395746	·0398278	− ·0002532	− ·6357
0 — − ·1	·0398121	·0398278	− ·0000157	− ·0394
− ·1 — − ·5	·1537784	·1516347	+ ·0021437	+ 1·4137
− ·5 — − 1·0	·1550725	·1498822	+ ·0051903	+ 3·4629
− 1·0 — − 1·5	·0952987	·0918481	+ ·0034506	+ 3·7569
− 1·5 — − 2·5	·0594246	·0605975	− ·0011729	− 1·9356

TABLE VIII.

Percentage Frequency Errors, when we use a Normal Curve having its Mean at the Median of u.

A. $u = kx^3$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0533032	·0605975	− ·0072943	−12·0373
1·5 — 1·0	·0935296	·0918481	+ ·0016815	+ 1·8307
1·0 — ·5	·1557551	·1498822	+ ·0058729	+ 3·9183
·5 — ·1	·1545682	·1516347	+ ·0029335	+ 1·9346
·1 — 0	·0398599	·0398278	+ ·0000321	+ ·0806
0 — − ·1	·0395432	·0398278	− ·0002846	− ·7146
− ·1 — − ·5	·1478606	·1516347	− ·0037741	− 2·4889
− ·5 — −1·0	·1434318	·1498822	− ·0064504	− 4·3036
−1·0 — −1·5	·0894952	·0918481	− ·0023529	− 2·5617
−1·5 — −2·5	·0674807	·0605975	+ ·0068832	+11·3589

B. $u = k\sqrt{x}$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0624532	·0605975	+ ·0018557	+3·0623
1·5 — 1·0	·0913867	·0918481	− ·0004614	− ·5024
1·0 — ·5	·1483234	·1498822	− ·0015588	−1·0400
·5 — ·1	·1507070	·1516347	− ·0009277	− ·6118
·1 — 0	·0397603	·0398278	− ·0000675	− ·1695
0 — − ·1	·0398397	·0398278	+ ·0000119	+ ·0299
− ·1 — − ·5	·1523872	·1516347	+ ·0007525	+ ·4963
− ·5 — −1·0	·1514045	·1498822	+ ·0015223	+1·0157
−1·0 — −1·5	·0923804	·0918481	+ ·0005323	+ ·5795
−1·5 — −2·5	·0588149	·0605975	− ·0017826	−2·9418

C. $u = \frac{k}{x}$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0678982	·0605975	+ ·0073007	+ 12·0479
1·5 — 1·0	·0898361	·0918481	− ·0020120	− 2·1906
1·0 — ·5	·1434150	·1498822	− ·0064672	− 4·3149
·5 — ·1	·1474715	·1516347	− ·0041632	− 2·7455
·1 — 0	·0394156	·0398278	− ·0004122	− 1·0350
0 — − ·1	·0397311	·0398278	− ·0000967	− ·2428
− ·1 — − ·5	·1541654	·1516347	+ ·0025307	+ 1·6689
− ·5 — −1·0	·1557627	·1498822	+ ·0058805	+ 3·9234
−1·0 — −1·5	·0939132	·0918481	+ ·0020651	+ 2·2484
−1·5 — −2·5	·0535366	·0605975	− ·0070609	− 11·6521

D. $u = k \log x$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0649345	·0605975	+ ·0043370	+7·1571
1·5 — 1·0	·0909481	·0918481	− ·0009000	− ·9799
1·0 — ·5	·1466458	·1498822	− ·0032364	−2·1593
·5 — ·1	·1495423	·1516347	− ·0020924	−1·3799
·1 — 0	·0396219	·0398278	− ·0002059	− ·5170
0 — − ·1	·0397802	·0398278	− ·0000476	− ·1195
− ·1 — − ·5	·1528967	·1516347	+ ·0012620	+ ·8323
− ·5 — −1·0	·1528144	·1498822	+ ·0029322	+1·9563
−1·0 — −1·5	·0929643	·0918481	+ ·0011162	+1·2153
−1·5 — −2·5	·0569577	·0605975	− ·0036398	−6·0065

however, is worthy of note. If we are dealing with numerical variations only, i.e. integrating u' from $-\chi\sigma_u$ to $+\chi\sigma_u$, then if terms in v_x^2 are negligible we have

$$\frac{1}{\sqrt{2\pi}}\int_{-\chi\sigma_u}^{+\chi\sigma_u} e^{-\frac{1}{2}\left(\frac{u'}{\sigma_u}\right)^2}\frac{du'}{\sigma_u} = \frac{1}{\sqrt{2\pi}}\int_{x_1'}^{x_2'} e^{-\frac{1}{2}\left(\frac{x'}{\sigma_x}\right)^2}\frac{dx'}{\sigma_x} \quad \text{............(38)},$$

or, a normal distribution does give the true frequency.

In view of the fact that we cannot obtain a normal curve for u for which the coefficient of v_x will vanish, it seemed worth while to investigate what improvement might result by shifting the origin of the normal curve for u. We have so far tested for the origin of the normal curve at the mean of u; other reasonable assumptions would be the origin at the median of u and at the mode of u, without in either case changing the constant σ_u.

Table VIII shows the percentage errors which arise when we shift the origin of the normal curve to the median of u, i.e. to the point which corresponds in u to $x'/\sigma_x = 0$. The general result of this shifting is to reduce in all cases the percentage errors of the central frequencies, + 1·5 to − 1·5, but the tail frequencies on either side are emphasised. In the case of kx^3 and k/x they are badly emphasised.

Lastly I considered what would happen, if the frequency curve for u' were treated as the normal curve shifted to the mode of u.

In order to deal with this, we must first note that the true frequency curve is

$$y = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_x} e^{-\frac{1}{2}\frac{(x-\bar{x})^2}{\sigma_x^2}} \quad \text{.........................(39)},$$

and accordingly the true form of the frequency curve for u is

$$y = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_x} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma_x}\right)^2}\frac{dx}{du} \quad \text{.........................(40)},$$

where x will be a function of u.

We will consider this curve for our special cases, determining the mode and modal ordinate.

A. $u = kx^3$.

Here $$y = \frac{1}{3\sqrt{2\pi}}\frac{1}{\sigma_x} k^{-\frac{1}{3}} u^{-\frac{2}{3}} e^{-\frac{1}{2}\left(\frac{\left(\frac{u}{k}\right)^{\frac{1}{3}} - \bar{x}}{\sigma_x}\right)^2}.$$

For our special case $k = 1$, $\sigma_x = 4$. Differentiating, the modal value of u is given by the equation

$$u_{mo}^{\frac{2}{3}} - u_{mo}^{\frac{1}{3}} + 32 = 0.$$

This leads to

$$u_{mo} = 990{,}399{\cdot}966, \quad u'_{mo} = -144{,}00{\cdot}033, \quad u'_{mo}/\sigma_u = -{\cdot}119{,}6179,$$

and the modal ordinate

$$y_{mo} = {\cdot}401{,}5067/\sigma_u.$$

B. $u = k\sqrt{x}$.

Here $$y = \sqrt{\frac{2}{\pi}}\,\frac{1}{\sigma_x}\,k^{-2}u\,e^{-\frac{1}{2}\left(\frac{\left(\frac{u}{k}\right)^2 - \bar{x}}{\sigma_x}\right)^2} \quad \text{....................}(41),$$

or, with the same values as before we have

$$u_{mo}{}^4 - 100u_{mo}{}^2 - 8 = 0,$$

leading to

$$u_{mo} = 10{\cdot}003{,}996066,\quad u'_{mo} = {\cdot}005{,}999076,\quad u'_{mo}/\sigma_u = {\cdot}029{,}974315,$$

and
$$y_{mo} = {\cdot}399{,}3024/\sigma_u.$$

C. $u = k/x$.

We have $$y = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma_x}\,ku^{-2}\,e^{-\frac{1}{2}\left(\frac{ku^{-1} - \bar{x}}{\sigma_x}\right)^2} \quad \text{...... ....}(42).$$

This results—with the same same values of σ_x, k and $\bar{x}$—in

$$32u_{mo}{}^2 + 100u_{mo} - 1 = 0,$$

or, $$u_{mo} = {\cdot}009{,}968203,\quad u'_{mo} = -{\cdot}000{,}04787423,\quad u'_{mo}/\sigma_u = -{\cdot}118{,}9165,$$

and
$$y_{mo} = {\cdot}4028054/\sigma_u.$$

D. $u = k \log x$.

Here $$y = \frac{1}{\sqrt{2\pi}}\,\frac{1}{\sigma_x}\,k^{-1}\,e^{u/k}\,e^{-\frac{1}{2}\left(\frac{e^{u/k} - \bar{x}}{\sigma_x}\right)^2}.$$

The equation for u_{mo} is

$$e^{2u_{mo}} - 100e^{u_{mo}} - 16 = 0,$$

leading to $e^{u_{mo}} = 100{\cdot}159{,}7448$, which gives $u_{mo} = 4{\cdot}606{,}766$, $u'_{mo} = {\cdot}002{,}398106$, $u'_{mo}/\sigma_u = {\cdot}059{,}76092$ and $y_{mo} = {\cdot}400{,}5420/\sigma_u$.

The y_{mo} have all been expressed in terms of σ_u as unit, so that in the diagrams u'/σ_u may be taken as abscissa and the curves still represent the frequency.

Table IX gives the percentage errors of the frequencies when we take the normal curve to have its origin at the mode of the u-distribution.

Examining Tables VII, VIII and IX we see that the best results between $x'/\sigma_x = -1{\cdot}0$ to $+1{\cdot}0$ are obtained by placing the origin of our normal curve at the mode of u. Between $x'/\sigma_x = -1{\cdot}5$ to about $-2{\cdot}0$ and between $x'/\sigma_x = 1{\cdot}5$ to $2{\cdot}0$, the normal curve with its origin at the median gives a lesser percentage error in estimating the frequency, while in the range outside the limits $x'/\sigma_x = \pm 2$ the normal curve with its origin at the mean of u is most adequate, although in the case of kx^3 and k/x it will still lead to errors in evaluating these tails of 4 to 5 %*.

In Diagrams I to IV the true curve of distribution of u'/σ_u is given for the four functions

$$u = kx^3,\quad u = k\sqrt{x},\quad u = k/x,\quad \text{and}\quad u = k\log x.$$

---

* **Table VI indicates that the round values here given for x'/σ_x may for practical purposes be taken to be those for u'/σ_u.**

TABLE IX.

Percentage Frequency Errors, when we use a Normal Curve having its Mean at the Mode of u.

A. $u = kx^3$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0454771	·0605975	− ·0151204	− 24·9522
1·5 — 1·0	·0841954	·0918481	− ·0076527	− 8·3319
1·0 — ·5	·1461984	·1498822	− ·0036838	− 2·4578
·5 — ·1	·1504410	·1516347	− ·0011937	− ·7872
·1 — 0	·0395754	·0398278	− ·0002524	− ·6337
0 — − ·1	·0395741	·0398278	− ·0002537	− ·6370
− ·1 — − ·5	·1508479	·1516347	− ·0007868	− ·5189
− ·5 — − 1·0	·1513121	·1498822	+ ·0014299	+ ·9540
− 1·0 — − 1·5	·0978835	·0918481	+ ·0060354	+ 6·5711
− 1·5 — − 2·5	·0772117	·0605975	+ ·0166142	+ 27·4173

B. $u = k\sqrt{x}$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0647514	·0605975	+ ·0041539	+ 6·8549
1·5 — 1·0	·0936016	·0918481	+ ·0017535	+ 1·9091
1·0 — ·5	·1504672	·1498822	+ ·0005850	+ ·3903
·5 — ·1	·1515662	·1516347	− ·0000685	− ·0452
·1 — 0	·0397921	·0398278	− ·0000357	− ·0896
0 — − ·1	·0397920	·0398278	− ·0000358	− ·0899
− ·1 — − ·5	·1514562	·1516347	− ·0001785	− ·1177
− ·5 — − 1·0	·1491555	·1498822	− ·0007267	− ·4848
− 1·0 — − 1·5	·0901062	·0918481	− ·0017419	− 1·8965
− 1·5 — − 2·5	·0566418	·0605975	− ·0039557	− 6·5278

C. $u = \frac{k}{x}$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0806815	·0605975	+ ·0200840	+ 33·1433
1·5 — 1·0	·0981614	·0918481	+ ·0063133	+ 6·8736
1·0 — ·5	·1512009	·1498822	+ ·0013187	+ ·8798
·5 — ·1	·1504164	·1516347	− ·0012183	− ·8034
·1 — 0	·0394468	·0398278	− ·0003810	− ·9566
0 — − ·1	·0394519	·0398278	− ·0003759	− ·9438
− ·1 — − ·5	·1501028	·1516347	− ·0015319	− 1·0103
− ·5 — − 1·0	·1463195	·1498822	− ·0035627	− 2·3770
− 1·0 — − 1·5	·0846402	·0918481	− ·0072079	− 7·8476
− 1·5 — − 2·5	·0457455	·0605975	− ·0148520	− 24·5093

D. $u = k \log x$.

$\frac{x'}{\sigma_x}$	N'	N	$\Delta \equiv N' - N$	$\frac{100\Delta}{N}$
2·5 — 1·5	·0691704	·0605975	+ ·0085729	+ 14·1473
1·5 — 1·0	·0952909	·0918481	+ ·0034428	+ 3·7484
1·0 — ·5	·1507969	·1498822	+ ·0009147	+ ·6103
·5 — ·1	·1511750	·1516347	− ·0004597	− ·3032
·1 — 0	·0396691	·0398278	− ·0001587	− ·3985
0 — − ·1	·0396699	·0398278	− ·0001479	− ·3713
− ·1 — − ·5	·1509796	·1516347	− ·0006551	− ·4320
− ·5 — − 1·0	·1482445	·1498822	− ·0016377	− 1·0927
− 1·0 — − 1·5	·0883841	·0918481	− ·0034640	− 3·7714
− 1·5 — − 2·5	·0530121	·0605975	− ·0075854	− 12·5177

DIAGRAM I

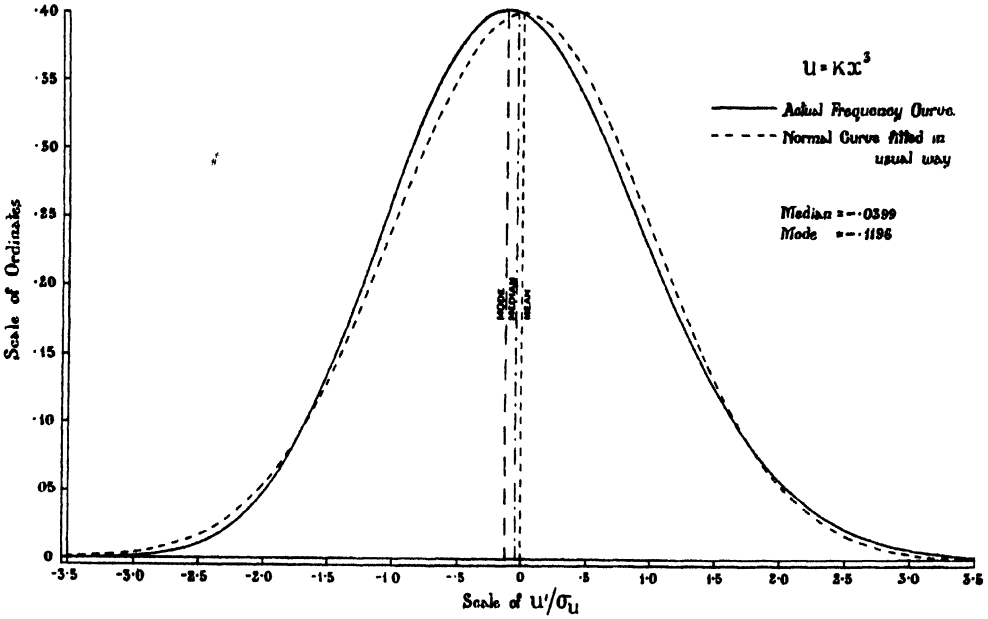


DIAGRAM II

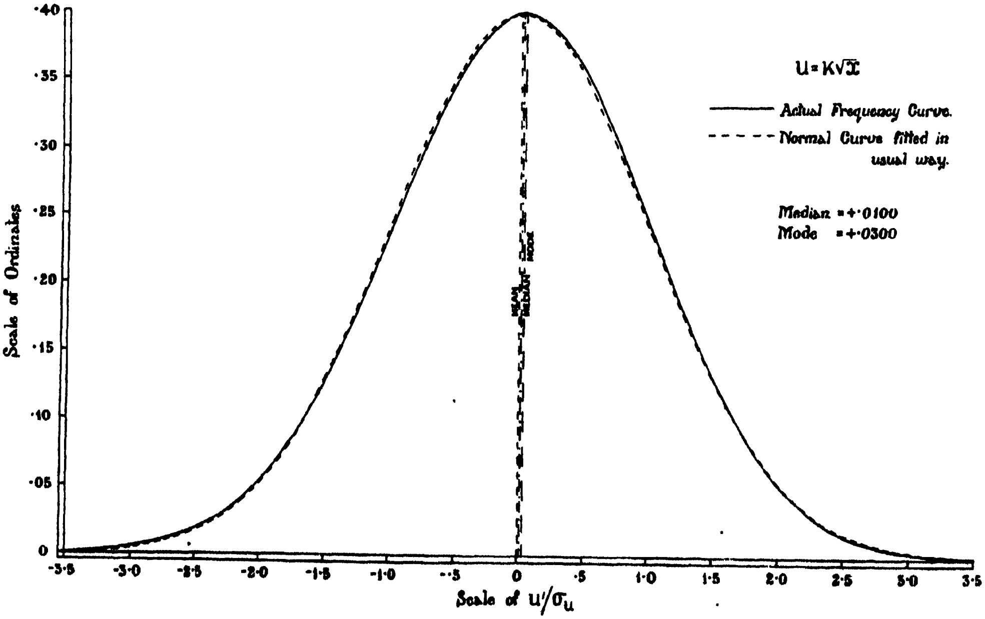


DIAGRAM III

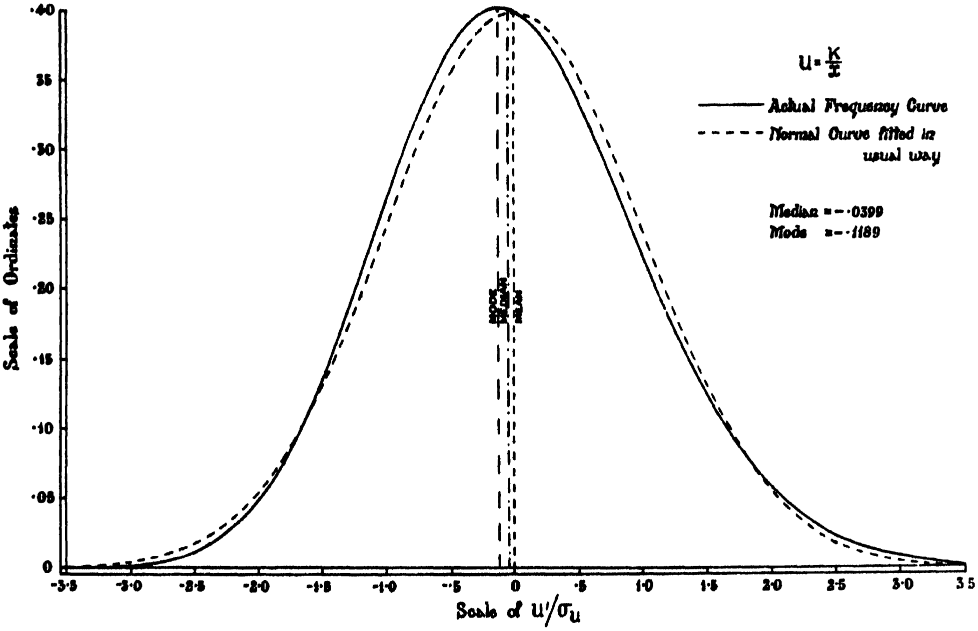
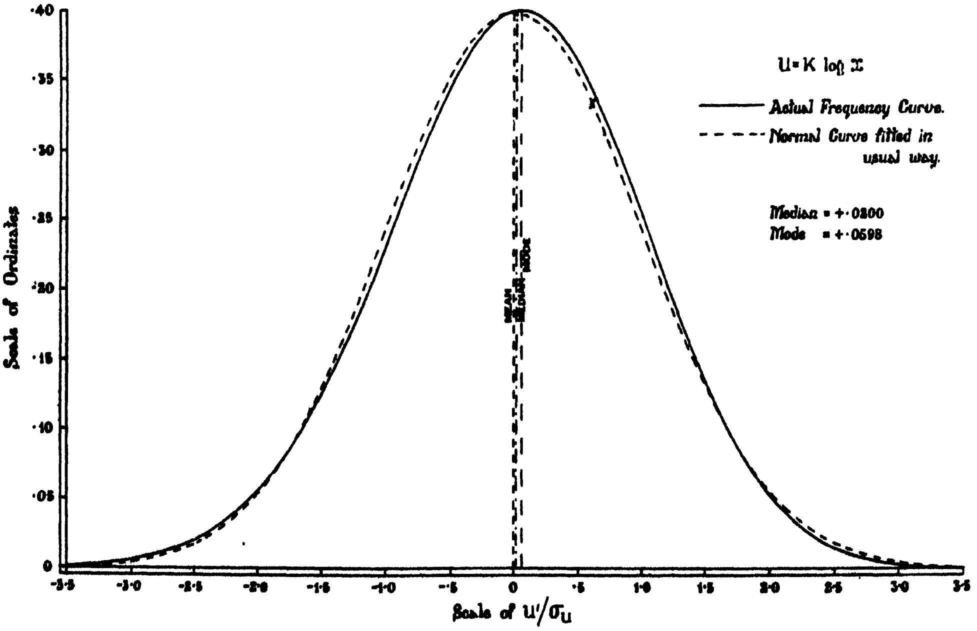


DIAGRAM IV



In each case the normal curve with origin at the mean and standard deviation σ_u is depicted. If the reader supposes this normal curve to be bodily shifted until its axis passes from the mean to the mode of u, he will readily understand how the fit is rendered worse at the tails, improves in the middle region when we reach the median and is best round the axis when we reach the mode.

(5) *General Conclusions.*

(*a*) Formulæ have been found enabling us to find the mean and standard deviation of a variable u which is a function of x in terms of those of x.

(*b*) It is shown that the correlation coefficient of u and x is for all cases likely to occur in ordinary practice (i.e. the customary values of v_x) very close to unity.

(*c*) Given two variates u and w, functions respectively of x and y, an expression is obtained for the correlation coefficient r_{uw}, and it is shown that for the customary values of v_x and v_y, r_{uw} is very nearly r_{xy}.

(*d*) Finally an investigation is made into the frequency distribution of u, if that of x is known to be normal. It is shown that a normal curve for u fitted in the usual way by the mean and standard deviation of u does not give very good results; errors may run up to 7 °/$_\circ$\*. Better agreement is obtained if we shift this normal curve bodily from the mean of the u-distribution through its median to the mode. It is then seen that the modal centering gives the best result when u'/σ_u lies between $-1{\cdot}0$ and $+1{\cdot}0$; that the median centering works best for the ranges $u'/\sigma_u = 1{\cdot}0$ to $2{\cdot}0$ or $-1{\cdot}0$ to $-2{\cdot}0$, while the mean centering gives the best results for the tails when u'/σ_u is greater than $2{\cdot}0$ numerically.

(*e*) If we desire a frequency stretching from $u'/\sigma_u = -\chi$ to $+\chi$, where χ has any numerical value, then the mean centered normal curve for u' will be found to give reasonable results owing to the vanishing of the odd terms multiplying v_x.

(*f*) It is always possible to calculate the exact value of the frequency of u, by determining from the function which u is of x, the values of x which correspond to the required values of u (as in Equations (26), (29), (32) and (35)) and then computing the known frequency of x, which will be that of the required range of u.

(*g*) In most statistical investigations, however, one variate is not any actual mathematical function of a second. Thus skull-capacity and brain-weight are not mathematically determined by the product of three diameters of the head, and if the latter three follow normal distributions, the former two are just as likely to do so also, as not to do so. For such physiological functional relations, the mathematical corresponding functions only roughly shadow reality, and even percentage errors of 7 or more are extremely likely to be swamped by the mass of additional variates on which the given variate really depends for its variation, besides those we endeavour to comprehend in the "mathematical" function.

I am indebted to Professor Pearson for his kindly advice and criticism during the course of this investigation, and to Miss Ida McLearn for the preparation of the diagrams.

---

* Perhaps 1 °/$_\circ$ of the *total* frequency.

ON A CRITERION FOR THE REJECTION OF OUTLYING OBSERVATIONS.

BY J. O. IRWIN, M.A., M.SC.

IN a previous paper[*] it was shown that if samples of a given size be taken at random from a normal population and if the individuals of each be arranged in descending order of magnitude with regard to some character, it is possible to obtain the frequency distribution of the differences between the pth and $p+1$th individuals in such samples.

For $p=1$ and for $p=2$, that is for the differences between the first and second and the differences between the second and third individuals, these frequency distributions were found to be very closely representable by curves of the form

$$y = y_0 e^{-\frac{1}{2}\left\{\frac{(x+h)^2-h^2}{\Sigma^2}\right\}},$$

the range being $x=0$ to $x=\infty$.

Diagrams were given from which it is possible to obtain the appropriate values of y_0, h and Σ for samples of any size up to 1000 individuals. We are thus able to calculate from these frequency distributions the probability that the difference (1) between the first and second and (2) between the second and third individuals in a sample of given size should be greater than λ times the standard-deviation of the population from which the sample is taken, where we may give λ any numerical value we please. If this probability becomes sufficiently small it clearly becomes admissible to reject (1) the first and (2) the first two individuals as not belonging to the same homogeneous group as the remainder. Thus a table of this probability for varying values of λ provides a criterion for the rejection[†] of outlying observations.

The frequency distribution of the differences between the pth and $p+1$th individuals ($p=1$ or 2) is given by

$$y = y_0 e^{-\frac{1}{2}\left\{\frac{(x+h)^2-h^2}{\Sigma^2}\right\}}.$$

---

[*] J. O. Irwin, "The Further Theory of Francis Galton's Individual Difference Problem," *Biometrika*, Vol. XVII. pp. 100—128.

[†] By rejection we mean the realisation of the fact that the particular observations in question probably do not belong to the same homogeneous group as the rest, and may therefore be left out of consideration in calculations concerning this group. If our observations are all observations of the same physical quantity we have a frequency distribution of errors and we realise that the outlying observations are due to some disturbing cause, possibly a blunder in measurement. But if the observations are a series of measurements of some character in different individuals, the outlying observations are those of individuals who are anomalous with regard to the character in question. We merely reject them because we realise they do not belong to the group we are studying; for other purposes such anomalies may be of the highest interest and importance.

Let the standard-deviation of the original population be σ, let $x=x'\Sigma$, $h=h'\Sigma$, $\lambda\sigma=\lambda'\Sigma$ and let $P(\lambda)$ be the chance of a difference being greater than $\lambda\sigma$.

Then
$$P(\lambda)=\frac{y_0\int_{\lambda\sigma}^{\infty} e^{-\frac{1}{2}\left\{\frac{(x+h)^2-h^2}{\Sigma^2}\right\}}\,dx}{y_0\int_{0}^{\infty} e^{-\frac{1}{2}\left\{\frac{(x+h)^2-h^2}{\Sigma^2}\right\}}\,dx}$$
$$=\frac{y_0\Sigma\int_{\lambda'}^{\infty} e^{-\frac{1}{2}\{(x'+h')^2-h'^2\}}\,dx'}{y_0\Sigma\int_{0}^{\infty} e^{-\frac{1}{2}\{(x'+h')^2-h'^2\}}\,dx'}$$
$$=\frac{\int_{\lambda'}^{\infty} e^{-\frac{1}{2}(x'+h')^2}\,dx'}{\int_{0}^{\infty} e^{-\frac{1}{2}(x'+h')^2}\,dx'}.$$

Let $x'+h'=X$.

Then
$$P(\lambda)=\frac{\int_{h'+\lambda'}^{\infty} e^{-\frac{1}{2}X^2}\,dX}{\int_{h'}^{\infty} e^{-\frac{1}{2}X^2}\,dX}=\frac{\frac{1}{2}(1-\alpha_{h'+\lambda'})}{\frac{1}{2}(1-\alpha_{h'})}.$$

Σ/σ and h/σ being known when the size of the sample is given, it is possible to compute tables of $P(\lambda)$ from the tables of the probability integral. Table I gives the values of h/σ and Σ/σ for values of n from 2 to 1000. These values have

TABLE I.

n	$p=1$		$p=2$	
	h/σ	Σ/σ	h/σ	Σ/σ
2	·00	1·414	—	—
3	·52	1·271	·52	1·271
10	1·29	1·095	1·61	·878
20	1·72	1·082	1·91	·822
30	1·91	1·072	2·06	·802
40	2·04	1·065	2·20	·793
50	2·13	1·060	2·30	·785
60	2·20	1·055	2·37	·782
70	2·26	1·051	2·43	·779
80	2·31	1·048	2·49	·777
90	2·35	1·047	2·54	·775
100	2·38	1·046	2·59	·774
200	2·63	1·038	2·80	·765
300	2·76	1·033	2·92	·758
400	2·84	1·030	3·01	·752
500	2·90	1·027	3·07	·748
600	2·95	1·026	3·11	·745
700	3·00	1·025	3·13	·742
800	3·04	1·025	3·15	·739
900	3·08	1·025	3·16	·736
1000	3·12	1·025	3·17	·733

been obtained from the diagrams published in the previous paper*, and by their aid Tables II and III have been computed. Table II gives $P_1(\lambda)$, the probability that the first and second individuals should differ by more than λ times the standard-deviation of the population from which the sample is drawn; Table III gives the same function ($P_2(\lambda)$) for the second and third individuals.

It is clear that Table II provides a criterion for the rejection of one outlying observation. It is a matter of opinion, which the mathematician cannot settle, how small $P(\lambda)$ must be before it is justifiable to reject the observation, but about ·05 seems reasonable while ·01 is perhaps on the side of caution. If the first observation has been rejected the second may be examined in the same way, using $(n-1)$ as the number of observations instead of n. Strictly λ is the ratio of the difference between the first and second individuals to the standard-deviation of the original population. The latter is in general unknown and we are obliged to take the standard-deviation of the sample as the best approximation to it obtainable. In calculating the standard-deviation of the sample we are on the safe side if we include the outlying observation.

If it happens that there are two outlying observations rather close together and then a large gap, we should turn to Table III and find $P_2(\lambda)$ the chance that the second and third individuals should differ by as much or more than they do. If we find $P_2(\lambda)$ is sufficiently small, say less than ·01, we are justified in rejecting the first two observations.

Suppose now that there are p observations ($p > 2$) close together and then a large gap, we really require the probability integral of the distribution of differences between the pth and $p+1$th individuals to deal with this case; but the numerical values of the constants of these distributions have not so far been determined. If however we remember that the probability of the pth and $p+1$th individuals differing by more than a given amount (in a sample of given size n) decreases with p† $\left(\text{until } p = \frac{n}{2} - 1 \text{ if } n \text{ is even and } \frac{n-1}{2} \text{ if } n \text{ is odd}\right)$ it follows that the value of $P(\lambda)$ obtained by using Table III when p is greater than 2 is really too large. Consequently if this value is less than say ·01, the true value will be still less and the outlying observations may be rejected. There may of course be some border line cases such as would arise, for example, if we had $p = 4$ and $P(\lambda) = ·1$ (as given by Table III). We should know that the true value of $P(\lambda)$ was less than ·1 but we should not know how much less. In cases of doubt retention rather than rejection would seem the safer course. As it does not often

---

* *Biometrika*, Vol. XVII. pp. 123—5.

† This has not been rigorously demonstrated. It follows easily from two assumptions.

(1) That the mean difference between pth and $p+1$th individuals in a sample of given size decreases as p increases until $p = \frac{n}{2} - 1$ if n be even and $\frac{n-1}{2}$ if n be odd.

(2) That the frequency distribution of differences between the pth and $p+1$th individuals is a J curve with range from 0 to ∞. The numerical evidence that we have supports these assumptions. See *Biometrika*, Vol. I. p. 397 and Vol. XVII. pp. 104—122.

TABLE II.

Values of $P_1(\lambda)$, or of the Probability that the First and Second Individuals in a Random Sample of n differ by more than λ times the Standard-Deviation of the Original Population.

	λ																								
n	·1	·2	·3	·4	·5	·6	·7	·8	·9	1·0	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0	2·1	2·2	2·3	2·4	2·5
2	·944	·888	·832	·777	·724	·671	·621	·572	·524	·480	·437	·396	·358	·322	·289	·258	·229	·203	179	·157	·138	·120	·104	·090	·077
3	·917	·836	·760	·687	·618	·553	·493	·437	·386	·339	·296	·257	·222	·191	·163	·139	·118	·099	·083	·069	·057	·047	·039	·031	·025
10	·855	·727	·613	·513	·427	·353	·289	·235	·190	·152	·121	·096	·075	·059	·045	·035	·026	·020	·015	·011	·008	·007	·004	·003	·002
20	·827	·679	·553	·447	·359	·285	·226	·177	·138	·107	·082	·062	·047	·035	·026	·019	·014	·010	·007	·005	·004	·003	·002	·001	·001
30	·813	·656	·525	·417	·328	·257	·199	·153	·117	·089	·068	·050	·037	·027	·020	·014	·010	·007	·005	·004	·003	·002	·001	·001	·001
40	·803	·639	·505	·396	·308	·238	·182	·138	·104	·078	·058	·042	·031	·022	·016	·011	·008	·006	·004	·003	·002	·001	·001	·001	
50	·796	·628	·491	·382	·294	·225	·170	·128	·096	·071	·052	·038	·027	·019	·014	·010	·007	·005	·003	·002	·001	·001	·001		
60	·790	·619	·481	·371	·283	·215	·161	·120	·089	·065	·048	·034	·025	·017	·012	·009	·006	·004	·003	·002	·001	·001	·001		
70	·785	·611	·471	·361	·274	·206	·154	·114	·084	·061	·044	·032	·022	·016	·011	·008	·005	·004	·002	·002	·001	·001			
80	·780	·604	·464	·353	·267	·200	·148	·109	·080	·058	·041	·030	·021	·015	·010	·007	·005	·003	·002	·001	·001	·001			
90	·778	·600	·459	·348	·262	·195	·144	·106	·077	·055	·040	·028	·020	·014	·010	·007	·004	·003	·002	·001	·001	·001			
100	·775	·596	·454	·343	·257	·191	·141	·103	·075	·054	·038	·027	·019	·013	·009	·006	·004	·003	·002	·001	·001	·001			
200	·757	·567	·422	·311	·227	·165	·118	·084	·060	·042	·029	·020	·014	·009	·006	·004	·003	·002	001	·001					
300	·746	·552	·405	·294	·212	·152	·107	·075	·052	·036	·025	·017	·011	·007	·005	·003	·002	·001	·001	·001					
400	·740	·543	·394	·284	·203	·144	·101	·070	·048	·033	·022	·015	·010	·007	·004	·003	·002	001	·001						
500	·735	·536	·387	·277	·197	·138	·096	·067	·045	·031	·021	·014	·009	·006	·004	·002	·002	·001	·001						
600	·731	·530	·381	·271	·191	·134	·093	·064	·043	·029	·020	·013	·009	·005	·004	·002	·001	·001	·001						
700	·727	·525	·375	·266	·187	·130	·090	·061	·041	·028	·019	·012	·008	·005	·003	·002	·001	·001	·001						
800	·725	·521	·371	·262	·183	·127	·087	·059	·040	·027	·018	·012	·008	·005	·003	·002	·001	·001							
900	·722	·517	·367	·258	·180	·124	·085	·058	·039	·026	·017	·011	·007	·005	·003	·002	·001	·001							
1000	·720	·514	·364	·255	·177	·122	·083	·056	·038	·025	·016	·011	·007	·005	·003	002	·001	·001							

	λ																								
n	2·6	2·7	2·8	2·9	3·0	3·1	3·2	3·3	3·4	3·5	3·6	3·7	3·8	3·9	4·0	4·1	4·2	4·3	4·4	4·5	4·6	4·7	4·8	4·9	5·0
2	·066	·056	·048	·040	·034	·028	·024	·020	·016	·013	·011	·009	·007	·006	·005	·004	·003	·002	·002	·001	·001	·001	·001	·001	—
3	·021	·016	·013	·010	·008	·006	·005	·004	·003	·002	·002	·001	·001	·001	·001										
10	·002	·001	·001	·001																					
20	·001																								

TABLE III.

Values of $P_2(\lambda)$, or of the Probability that the Second and Third Individuals in a Random Sample of n differ by more than λ times the Standard-Deviation of the Original Population.

	λ																			
n	·1	·2	·3	·4	·5	·6	·7	·8	·9	1·0	1·1	1·2	1·3	1·4	1·5	1·6	1·7	1·8	1·9	2·0
3	·917	·836	·760	·687	·618	·553	·493	·437	·386	·339	·296	·257	·222	·191	·163	·139	·118	·099	·083	·069
10	·772	·589	·444	·331	·244	·177	·128	·091	·064	·044	·030	·021	·014	·009	·006	·004	·002	·002	·001	
20	·718	·509	·356	·246	·167	·112	·074	·049	·031	·020	·012	·008	·005	·003	·002	·001	·001			
30	·693	·473	·319	·211	·138	·089	·057	·035	·022	·013	·008	·004	·003	·001	·001					
40	·674	·447	·293	·189	·120	·075	·046	·028	·017	·010	·006	·003	·002	·001	·001					
50	·659	·428	·274	·172	·107	·065	·039	·023	·014	·008	·004	·002	·001	·001						
60	·650	·416	·262	·163	·099	·060	·035	·021	·012	·007	·004	·002	·001	·001						
70	·642	·406	·252	·155	·093	·056	·032	·019	·011	·006	·003	·002	·001							
80	·635	·397	·244	·148	·088	·052	·029	·017	·010	·005	·003	·002	·001							
90	·628	·388	·237	·142	·084	·049	·027	·016	·009	·005	·003	·001	·001							
100	·623	·382	·230	·137	·080	·046	·026	·015	·008	·004	·002	·001	·001							
200	·596	·349	·201	·114	·064	·035	·019	·010	·005	·003	·001	·001								
300	·579	·329	·184	·101	·055	·029	·015	·008	·004	·002	·001									
400	·565	·314	·172	·092	·049	·025	·013	·006	·003	·002	·001									
500	·557	·305	·164	·086	·045	·023	·012	·006	·003	·001	·001									
600	·550	·297	·158	·082	·042	·021	·011	·005	·002	·001	·001									
700	·545	·292	·154	·080	·041	·020	·010	·005	·002	·001										
800	·541	·288	·150	·077	·039	·019	·009	·005	·002	·001										
900	·537	·284	·147	·075	·038	·018	·009	·004	·002	·001										
1000	·533	·280	·144	·073	·036	·017	·008	·004	·002	·001										

	λ																			
n	2·1	2·2	2·3	2·4	2·5	2·6	2·7	2·8	2·9	3·0	3·1	3·2	3·3	3·4	3·5	3·6	3·7	3·8	3·9	4·0
3	·057	·047	·039	·031	·025	·020	·016	·013	·010	·008	·006	·005	·004	·003	·002	·002	·001	·001	·001	·001

happen that there are more than three or four outlying observations the above tables would seem to be adequate to deal with most cases that arise.

It seems desirable to mention one or two other criteria which have been suggested for the rejection of outlying observations and to compare the results they give with those given by the one suggested here.

(1) Peirce's Criterion* gives a complicated formula based on the principle "that the proposed observations should be rejected when the probability of the system of errors obtained by retaining them is less than that of the system obtained by their rejection multiplied by the probability of obtaining so many and no more abnormal observations." Apart from the artificiality of the principle the proof of Peirce's criterion reproduced in Chauvenet's astronomy "nearly in the words of its author and with only some slight changes in notation" seems in several places open to criticism. It is well-known to have been severely criticised by Airy†.

(2) Chauvenet gives a criterion for the rejection of one observation‡. Pointing out that the probability of an error less than t' times the probable error of a single observation is

$$\Theta(\rho t') = \frac{2}{\sqrt{\pi}} \int_0^{\rho t'} e^{-t^2}\, dt,$$

where

$$\rho = \frac{\cdot 67449}{\sqrt{2}} = \cdot 47694,$$

he shows that it follows that in m observations the number to be expected numerically greater than t' times the probable error will be

$$m\,[1 - \Theta(\rho t')],$$

or in modern notation

$$m\left[1 - \frac{2}{\sqrt{2\pi}\sigma} \int_0^{\cdot 67449 t'\sigma} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}\, dx\right].$$

Chauvenet then says "If this quantity is less than $\frac{1}{2}$, it will follow that an error of the magnitude rt' (r being the probable error) will have a greater probability against it than for it, and may therefore be rejected." He therefore rejects those errors for which

$$m\,[1 - \Theta(\rho t')] < \tfrac{1}{2},$$

or

$$\Theta(\rho t') > \frac{2m - 1}{2m}.$$

But the above argument is clearly wrong for $m\,[1 - \Theta(\rho t')]$ is not a probability but a number, e.g. if $m = 100$, $\Theta(\rho t') = \frac{1}{2}$,

$$m\,[1 - \Theta(\rho t')] = 50.$$

What Chauvenet really does is to reject errors greater than t' times the probable error where t' is such that only half an error is to be expected greater than this limit. But the choice of one half in this connection rather than any other proper fraction seems quite arbitrary. If one observation has been rejected by Chauvenet's criterion it may, he says, be applied to the next giving m a value less by one than before.

---

* W. Chauvenet, *A Manual of Practical and Spherical Astronomy*, Fourth Edition, Vol. II. p. 558.

† *Cambridge Astronomical Journal*, Vol. IV. p. 137.

‡ W. Chauvenet, *A Manual of Spherical and Practical Astronomy*, Fourth Edition, Vol. II. p. 565.

(3) A simple method that is often employed is to find from the ordinary tables of the probability integral the probability of an observation so divergent as the outlying one occurring at all. For example suppose we find in a series of 1000 observations one which is greater than the mean by 3·5 times the standard-deviation and that the one before it is greater than the mean by 3·0 times the standard-deviation. We find from the tables of the probability integral the chance of an observation occurring more distant from the mean than the mid-point between these two:—that is 3·25 times the standard-deviation from the mean. That is ·0006, or in 1000 observations we should expect ·6 of an observation to be so distant; and we should not be justified in rejecting it. But if the outlying observation were four times the standard-deviation from the mean, noting that the probability of a deviation greater than $3{\cdot}5\sigma$ is ·0002 and greater than $3{\cdot}75\sigma$ is ·0001, we should incline to rejection, while we should certainly reject an observation whose deviation from the mean was $4{\cdot}5\sigma$ seeing that the probability of a deviation greater than $4{\cdot}0\sigma$ is ·00003.

This test may well be used to supplement the criterion we have suggested in this paper. All these criteria are of course based on normality, they would not be applicable where the material is very different from the normal; but they will apply in most cases to errors of observation and to anthropometric measurements.

It must always be remembered that caution is necessary in dealing with small samples as the standard-deviation of the sample is then an unreliable measure of the standard-deviation of the original population.

We now proceed to discuss some examples:

(*a*) We take from Chauvenet's *Astronomy*[*] the following fifteen observations of the vertical semi-diameters of Venus made by Lieut. Herndon in 1846.

Deviations from the Mean.

− 0″·30	− 0″·24	− 1″·40	0″·18
− 0 ·44	+ 0 ·06	− 0 ·22	0 ·39
1 ·01	+ 0 ·63	− 0 ·05	+ 0 ·10
0 ·48	− 0 ·13	+ 0 ·20	

Peirce's criterion is applied by Chauvenet and leads to the rejection of two observations, − 1″·40 and 1″·01.

We have

$$\sigma = 0''{\cdot}5326,$$

$$\frac{2m-1}{2m} = \frac{29}{30} = {\cdot}9667.$$

---

[*] Fourth Edition, Vol. II. p. 562.

Thus Chauvenet's criterion would lead to the rejection of observations numerically greater than $k\sigma$, where

$$\frac{2}{\sqrt{2\pi}\sigma}\int_0^{k\sigma} e^{-\frac{1}{2}x^2/\sigma^2}\,dx = \cdot 9667,$$

or if, as usual,

$$\tfrac{1}{2}(1+\alpha_k) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{k\sigma} e^{-\frac{1}{2}x^2/\sigma^2}\,dx,$$

we have

$$\begin{aligned} \alpha_k &= \cdot 9667, \\ \tfrac{1}{2}(1+\alpha_k) &= \cdot 9833, \\ k\sigma &= 2{\cdot}13\sigma \\ &= 1''{\cdot}13. \end{aligned}$$

This leads to the rejection of $-1''{\cdot}40$.

For the remaining 14 observations

$$\sigma = 0''{\cdot}4048,$$

$$\frac{2m-1}{2m} = \frac{27}{28} = \cdot 9643.$$

Or we have

$$\begin{aligned} \alpha_k &= \cdot 9643, \\ \tfrac{1}{2}(1+\alpha_k) &= \cdot 9821, \\ k\sigma &= 2{\cdot}10\sigma = 0''{\cdot}8501. \end{aligned}$$

This leads to the rejection of $1''{\cdot}01$.

On the repetition of the process it is found that no more observations can be rejected.

Now let us apply our own criterion.

Difference between first and second

$$= 1''{\cdot}01 - 0''{\cdot}63 = 0''{\cdot}38,$$

$$\frac{\text{Difference}}{\sigma} = \frac{\cdot 38}{\cdot 5326} = \cdot 713.$$

Table II gives $P_1 = \cdot 241$.

Difference between last and last but one

$$= 1''{\cdot}40 - 0''{\cdot}44 = 0''{\cdot}96,$$

$$\frac{\text{Difference}}{\sigma} = \frac{\cdot 96}{\cdot 5326} = 1{\cdot}802.$$

Table II gives $P_1 = \cdot 014$.

This seems to indicate that $-1''{\cdot}40$ but not $1''{\cdot}01$ should be rejected.

We may note that

$$\frac{\frac{1}{2}(1''{\cdot}01 + 0''{\cdot}63)}{\sigma} = \frac{0''{\cdot}82}{\sigma} = \frac{\cdot 82}{\cdot 5326} = 1{\cdot}54 \qquad \tfrac{1}{2}(1-\alpha) = \cdot 062,$$

$$\frac{\frac{1}{2}(1''{\cdot}40 + 0''{\cdot}44)}{\sigma} = \frac{0''{\cdot}92}{\sigma} = \frac{\cdot 92}{\cdot 5326} = 1{\cdot}73 \qquad \tfrac{1}{2}(1-\alpha) = \cdot 042,$$

or in 15 observations we should expect

$$\begin{aligned} &\cdot 93 \text{ observations} > 0''\!\cdot\!82, \\ &\cdot 63 \quad\;\; ,, \qquad < -0''\!\cdot\!92. \end{aligned}$$

These results would not by themselves justify the rejection of either doubtful observation but, as we have seen, the greatness of the gap between $-1''\cdot40$ and the next observation would seem to lead to its rejection.

The fact that these tests would lead to the retention while Peirce's and Chauvenet's criteria would lead to the rejection of the observation $1''\cdot01$ probably indicates that the latter reject a little too easily. We must always remember however that in dealing with such a sample as 15, the standard-deviation of the sample is a very unreliable measure of the standard-deviation of the original population.

In fact the distribution of standard-deviations in samples of 15 taken at random from a normal population is given by the curve

$$y = y_0 x^{13} e^{-\frac{15x^2}{2\sigma^2}} *,$$

σ being the standard-deviation of the original population and the range being from $x = 0$ to $x = \infty$.

Now it is easily deduced from this curve that the probability of a standard-deviation in a sample being between 0 and x is

$$\frac{\Gamma_X(7)}{\Gamma(7)}, \text{ where } X = \frac{15x^2}{2\sigma^2},$$

and $\Gamma_X(7)$ is the incomplete Γ function $\int_0^X e^{-X} X^6\, dX$.

From the *Tables of the Incomplete Γ-function* it is now easily deduced that there is a probability of ·95 of the standard-deviation of a sample being greater than $\cdot664\sigma$ and likewise a probability of ·95 of it being less than $1\cdot25\sigma$ while the probability of it being less than σ is ·6220†, so that the unknown standard-deviation of the original population may well vary between 1·5 times and 0·8 times the standard-deviation of the sample.

If we had taken the former figure we should have found $P_1 = \cdot075$ for the observation $-1''\cdot40$ which would have brought it into the region where rejection is doubtful. If we had taken the latter figure in testing the observation $1''\cdot01$, we should have found $P_1 = \cdot162$ which would still not justify us in rejecting it.

In general if the standard-deviation of the sample be too low we may well reject observations we ought to retain, if it is too high the opposite will be the case; it is in fact more likely to be too low than too high, so that caution is required in rejecting observations from a small series. At the same time an astronomer long practised in observing has a wider experience of variation than that afforded by an individual sample of this kind. His opinion on the question of the rejection or retention of a doubtful observation in astronomical work should therefore carry some weight.

---

* *Biometrika*, Vol. VI. p. 10.

† The limits $\cdot664\sigma$ and $1\cdot25\sigma$ are not very different from those which would have been obtained by using the ordinary formula $\cdot67449\sigma/\sqrt{2n}$ for the probable error and taking $2\frac{1}{2}$ times the probable error on either side. For we have $\sigma = \cdot5826 \pm \cdot0656$ and the corresponding limits are $\cdot69\sigma$ and $1\cdot31\sigma$.

(β) Let us take the following capacities of 17 male Moriori Crania*.

Capacities in cubic centimetres.

1230	1360	1380	1445	1630
1260	1364	1410	1470	
1318	1378	1410	1540	
1348	1380	1420	1545	

We have Mean Capacity = 1405·2,

$$\sigma = 84·1 \pm 9·7 †.$$

Let us test 1630‡

$$\frac{1630 - 1545}{84·1} = \frac{85}{84·1} = 1·011.$$

From Table II taking $n = 17$ we find $P_1 = ·114$. This does not justify the rejection of the observation 1630. Taking the probable error into consideration in about 90 °/₀ of cases the standard-deviation of the original population will lie between 59·9 and 108·3. If we take the lower limit 59·9 we have

$$\frac{85}{59·9} = 1·419, \quad P_1 = ·037,$$

a value which is on the border line for rejection; but the value 59·9 being improbable, even taking into consideration only the information afforded by the sample, we are justified in concluding that the observation must be retained.

Using Chauvenet's criterion we have:

$$\alpha_k = \frac{2m-1}{2m} = \frac{33}{34} = ·97059,$$

$$\tfrac{1}{2}(1 + \alpha_k) = ·98529,$$

$$k\sigma = 2·178\sigma = 183·2.$$

Accordingly this would lead to the rejection of observations

$$> 1405·2 + 183·2 = 1588·4,$$

$$< 1405·2 - 183·2 = 1222·0.$$

Thus Chauvenet's criterion would lead to the rejection of 1630 which confirms our conclusion that it rejects somewhat too easily.

---

[*] *Biometrika*, Vol. XI. p. 134.

[†] Using the ordinary formula for the probable error which, as we have seen in the last example, is accurate enough for our purpose.

[‡] We have chosen this example not because we think a craniologist would think of rejecting the observation 1630 but to show (i) that the suggested criterion gives a result in accordance with the usual practice of the craniologist and (ii) that Chauvenet's criterion rejects somewhat too easily.

Using method (3) we have $\frac{1}{2}(1630 + 1545) = 1587{\cdot}5$,

$$\frac{1587{\cdot}5 - \text{mean}}{\sigma} = \frac{1587{\cdot}5 - 1405{\cdot}2}{84{\cdot}1} = 2{\cdot}17,$$

whence

$$\tfrac{1}{2}(1-\alpha) = {\cdot}015,$$

$$17\,({\cdot}015) = {\cdot}255,$$

which does not justify rejection.

(γ) As our last example we take 423 observations made by Dr R. A. Houston on the Colour Vision of male students, the Rayleigh test being used*. The variate is here the logarithm of the ratio of the intensity of red to green which appeared to the subject to be necessary in order that the combined colour might match a given yellow.

Logarithm of Ratio of Red to Green (x)	Frequency	Logarithm of Ratio of Red to Green (x)	Frequency
− 2·93†	1	− ·33	9
1·78	1	− ·28	37
− 1·68	1	·23	77
− 1·63	1	− ·18	112
− 1·38	1	− ·13	94
		− ·08	38
− ·73	3	− ·03	14
− ·68	3	·02	8
− ·63	3	·07	2
− ·58	5	·12	1
− ·53	2	·17	1
− ·48	4		
− ·43	0	·62	1
− ·38	2	·87	1
		1·07	1
		Total	423

There are five cases at one end of the scale and three at the other which appear to be anomalous; let us test these. We find

$$\bar{x} = -{\cdot}20305,$$

$$\sigma = {\cdot}24316.$$

Taking the difference between the observation ·62 and the preceding one and dividing by the standard-deviation we have

$$\frac{{\cdot}62 - {\cdot}17}{{\cdot}24316} = \frac{{\cdot}45}{{\cdot}24316} = 1{\cdot}851.$$

We have not a table of probabilities for differences between third and fourth individuals, but we know that the values given by Table III will be too large. A glance at this table is sufficient to assure us accordingly that the probability of such a difference arising purely by chance in a sample from a normal population is so small that it falls outside our table, that is to say it is less than ·0005. Thus the observations ·62, ·87 and 1·07 would be rejected and we can conclude that these three people have some peculiar anomaly of vision.

---

* *Proc. R. Soc.* 102 A, 1922–23, p. 353.

† I.e. $\geqslant(-2{\cdot}95)$ and $<(-2{\cdot}90)$.

To test the observation − 1·38 we have

$$\frac{1{\cdot}38 - {\cdot}73}{\sigma} = \frac{{\cdot}65}{{\cdot}24316} = 2{\cdot}673,$$

so that the same conclusion applies *a fortiori* to the five observations at the beginning of the data. There is no doubt that all the eight outlying observations are here anomalies.

Chauvenet's criterion leads to the same result.

We have
$$\alpha_k = \frac{2m-1}{2m} = \frac{845}{846} = {\cdot}99882,$$
$$\tfrac{1}{2}(1+\alpha_k) = {\cdot}99941,$$
$$k\sigma = 3{\cdot}24\sigma = {\cdot}788.$$

Hence we must reject observations

$$> (-{\cdot}203 + {\cdot}788) = {\cdot}585,$$
$$< (-{\cdot}203 - {\cdot}788) = -{\cdot}991,$$

which leads to the same conclusion.

Method (3) gives the following values of $\frac{1}{2}(1-\alpha)$.

Observations	$\frac{\text{Deviation from Mean}}{\sigma}$	$\frac{1}{2}(1-\alpha)$
−2·93	−11·23	$1{\cdot}46 \times 10^{-20}$
−1·78	− 6·50	$4{\cdot}00 \times 10^{-11}$
−1·68	− 6·09	$5{\cdot}64 \times 10^{-10}$
−1·63	− 5·88	$2{\cdot}05 \times 10^{-9}$
−1·38	− 4·85	$6{\cdot}17 \times 10^{-7}$
+ ·62	3·37	$3{\cdot}76 \times 10^{-4}$
·87	4·40	$5{\cdot}41 \times 10^{-6}$
1·07	5·22	$8{\cdot}95 \times 10^{-8}$

It is clear that we must reject all these observations except − 1·38 and + ·62 which perhaps require further testing.

Taking the next observation into consideration we have

$$\tfrac{1}{2}(-1{\cdot}38 - {\cdot}73) = -1{\cdot}055,$$
$$\frac{\text{Deviation from Mean}}{\sigma} = \frac{{\cdot}852}{\sigma} = 3{\cdot}50,$$
$$\tfrac{1}{2}(1-\alpha) = {\cdot}00023,$$

and the observation must be rejected.

To test + ·62 we have

$$\tfrac{1}{2}(\cdot 17 + \cdot 62) = \cdot 395,$$

$$\frac{\cdot 395 - (- \cdot 203)}{\sigma} = \frac{\cdot 598}{\sigma} = 2\cdot 46,$$

$$\tfrac{1}{2}(1 - \alpha) = \cdot 007,$$

or in 423 observations we should expect three greater than ·395. How many should we expect greater than ·5?

$$\frac{\cdot 5 - (- \cdot 203)}{\sigma} = \frac{\cdot 703}{\sigma} = 2\cdot 89,$$

$$\tfrac{1}{2}(1 - \alpha) = \cdot 002,$$

or we should expect ·8 observations > ·5,

and similarly ·4 „ > ·55.

Thus this test considered by itself might lead to the retention of the observation ·62 if we did not take into consideration the great gap between it and the previous one, ·17. But, allowing for the gap, it should, we conclude, be rejected.

NOTES ON AN EXPERIMENTAL TEST OF ERRORS IN PARTIAL CORRELATION COEFFICIENTS, DERIVED FROM FOURFOLD AND BISERIAL TOTAL COEFFICIENTS.

BY ETHEL M. NEWBOLD.

I.

IN many biometric and social investigations we are obliged to deal with variables which cannot be expressed quantitatively, and yet nevertheless partial correlation is necessary to unravel the skein of their interrelations. The total correlation coefficients in these cases can only be found by using biserial, tetrachoric or other fourfold *r*'s or coefficients of contingency, so that the question arises, how far are we justified in drawing conclusions from partial coefficients based on total coefficients of this kind? The present note describes an experimental test of this question, which was made for the Committee of Industrial Health Statistics of the Medical Research Council. The greater part of the arithmetical computation has been done by my fellow workers on the staff of that Committee—Mr E. Lewis-Faning, Mr J. Martin and Miss C. Thomas—and our thanks are due to Dr Major Greenwood and Dr L. Isserlis for helpful suggestions and advice.

In these tests we have confined ourselves chiefly to tetrachoric *r*'s but have also included some biserials. As regards the values of total tetrachoric coefficients, tests have previously been made on approximately normal distributions by Professor Pearson* and also by the late W. R. Macdonell†. The agreement with product moment values was good. Professor Pearson found also that the probable error of tetrachoric *r* increases with the distance of the dividing lines from the mean, but not very rapidly, and that the probable error of the tetrachoric *r* is 1·5 to 2 times the probable error as found by the product moment method, and also that it is not necessarily the case, in various positions of the dividing lines, that the smaller the probable error, the more nearly will the tetrachoric agree with the product moment coefficient. Deviations between the tetrachoric and product moment values arise from want of normality and not from errors of sampling (except in so far as these may affect the normality of the sample), hence the question of a suitable criterion for judging the reliability of these partial correlations needs consideration.

It seems clear first that, since the ordinary method of partial correlation is equivalent to correlating deviations from regression straight lines, and so involves total coefficients found by the product moment method, we must *for our present purpose* regard our coefficients as approximations to the product moment values, without in any way disparaging the claim that any total fourfold or other r may have to be considered as in itself a reasonable scale of association; such as, for example, Professor Pearson has suggested for his fourfold r deduced from "equality of improbability*," and also for his Q_5 coefficient†.

---

* *Phil. Trans.* A, 195, pp. 1—47.

† *Biometrika*, Vol. I. pp. 177—227.

We have, therefore, in these tests found the differences between the partial r's derived from totals calculated by tetrachoric or other methods and the product moment partials. We might then have compared these differences to their probable errors (or, rather, to approximate expressions for the latter, since they involve the unknown, but presumably high correlation in errors between product moment and tetrachoric r's derived from the same sample). This criterion does not however, for reasons given below, seem to give the most direct answer to the practical point at issue. A rather simpler way of putting the question is to ask whether an observer A, who bases his judgment on the product moment partial and its probable error, is likely to come to a different practical conclusion from B, who bases his judgment on the tetrachoric partial and its probable error. Since, however, the deviations between the two kinds of partials from the same sample depend primarily on the departure from normality in the distributions and not on sampling errors, and cannot be got rid of in a skew distribution by any multiplication of observations, it is, perhaps, more useful to consider the deviations observed in these tests either absolutely or relatively to the value of r, quite apart altogether from the probable errors of the particular data used here. Then, so far as the results may be regarded as typical of the sort of distributions examined, these tests will serve to give some idea of the order of the average differences between partials from the different methods to be expected in such distributions, i.e. the average displacement of the mean value about which the sampling errors fluctuate. Whether or not this displacement is likely to be of practical importance in any proposed case is a question to which a general answer cannot be given as it depends on the sampling error of the tetrachoric partial found in that particular case, and the probable errors of the values in our tests, depending as they do on the size of the particular sets of data chosen, are quite irrelevant. The probable errors given in the tables, therefore, are only meant as illustrations of particular cases and do not affect any general conclusions.

In a practical case, of course, since the tetrachoric would not be used if a quantitative classification were available, we have no means of knowing the distribution and may be dealing with a very skew surface; hence in our choice of material we have purposely included some variables which are very far from fulfilling the assumptions made in the approximate methods, and in one case give very artificial distributions, so that the deviations found in that case may be looked on as extreme. The other source of large deviations—extreme divisions—can always be recognised, hence as Professor Pearson has already illustrated it* we have avoided this as far as possible, and taken the dividing lines in every case as near the median as the data allowed.

---

* "On a Novel Method of regarding the Association of two Variates classed solely in alternative Categories." Math. Cont. to the Theory of Evolution XVIII. *Drapers' Co. Research Memoirs*, Biom. Series, VII. 1912, p. 4.

† "On the Correlation of Characters not quantitatively measurable." Math. Cont. to the Theory of Evolution VII. *Phil. Trans.* A, 195, 1900, pp. 17 and 18.

II. DESCRIPTION OF TESTS.

Three sets of data were used for the experimental tests.

Test I. The material chosen for the first test (taken from "A Biometric study of the Inter-relations of 'Vital Capacity,' Stature, Stem Length and Weight in a sample of healthy male adults" by Cripps, Greenwood and Newbold, *Biometrika*, Vol. XIV. p. 327), practically fulfils the assumptions made in both the approximate methods, as the approach to normal distributions in all the variables is fairly close (with the exception of a slight skewness in weight, and to a lesser degree in chest), and the deviations from linearity are small (see p. 327, *loc. cit.*). The position of the points of division also was taken very near the mean in each case, and the number of observations (950) is respectable, so that the conditions are about as favourable for the tetrachoric and biserial methods as they are likely to be in any practical case involving five variables.

Table I gives the results in detail and the values of β_1 and β_2 for the five variables. The average absolute differences are:

		Tetrachoric	Biserial	Tetrachoric and Biserial	Tetrachoric and Product Moment	Biserial and Product Moment
	Totals	·035	·014	—	—	—
Partials	1st order ...	·056	·019	·024	·018	·013
	2nd order ...	·071	·021	·035	·025	·015
	3rd order ...	·079	·047	·032	·027	·024

The largest deviation is $-·175$ for the tetrachorics and ·06 for the biserial r's, both in the 3rd order coefficients.

There seems to be no very definite tendency in any of these methods to deviate from the equal distributions of positive and negative differences, and the final order of accuracy of the methods is as would be expected:

(1) Biserial and Product Moment.
(2) Biserial alone.
(3) Tetrachoric and Product Moment.
(4) Tetrachoric and Biserial.
(5) Tetrachoric.

The deviations of the biserial from the product moment are on an average rather less than half those of the tetrachoric.

---

* *loc. cit.* p. 42.

X_1 = Chest, X_2 = Vital Capacity, X_3 = Weight, X_4 = Standing Height, X_5 = Sitting Height.
Number of Observations = 950.

Subscripts of *r*	Product Moment (*a*)	Tetrachoric (*b*)	Biserial (*c*)	Tetrachoric − Product Moment (*b*) − (*a*)	Biserial − Product Moment (*c*) − (*a*)	Tetrachoric and Biserial (*d*)	Difference between (*a*) and (*d*)	Tetrachoric and Product Moment (*e*)	Difference between (*a*) and (*e*)	Biserial and Product Moment (*f*)	Difference between (*a*) and (*f*)
12	·395 ± ·019	·415 ± ·030	·375	·020	− ·019	—	—	—	—	—	—
13	·741 ± ·010	·731 ± ·021	·746	− ·010	·005	—	—	—	—	—	—
14	·336 ± ·019	·264 ± ·033	·317	− ·072	− ·019	—	—	—	—	—	—
15	·367 ± ·019	·298 ± ·033	·349	− ·069	− ·019	—	—	—	—	—	—
23	·495 ± ·017	·546 ± ·027	·510	·052	·015	—	—	—	—	—	—
24	·586 ± ·014	·586 ± ·026	·592	·000	·006	—	—	—	—	—	—
25	·549 ± ·015	·493 ± ·028	·535	− ·056	− ·014	—	—	—	—	—	—
34	·544 ± ·015	·536 ± ·028	·532	− ·008	− ·012	—	—	—	—	—	—
35	·540 ± ·016	·541 ± ·028	·534	·001	− ·006	—	—	—	—	—	—
45	·720 ± ·011	·778 ± ·018	·747	·057	·026	—	—	—	—	—	—
21.5	·248 ± ·021	·322 ± ·039*	·238	·074	− ·010	·288	·040	·274	·026	·263	·015
24.5	·328 ± ·020	·371 ± ·038	·343	·043	·014	·320	− ·008	·303	− ·026	·342	·014
14.5	·111 ± ·022	·054 ± ·044	·092	− ·057	− ·019	·079	− ·032	·086	− ·025	·102	− ·009
23.5	·282 ± ·020	·382 ± ·037	·313	·100	·032	·310	·028	·281	− ·001	·310	·028
34.5	·266 ± ·020	·219 ± ·042	·238	− ·048	− ·029	·211	− ·055	·234	− ·033	·246	− ·021
13.5	·693 ± ·011	·709 ± ·022	·706	·017	·014	·688	− ·005	·680	− ·013	·700	·008
24.1	·524 ± ·016	·543 ± ·031	·538	·019	·014	·527	·003	·521	− ·003	·529	·005
23.1	·328 ± ·020	·392 ± ·037	·372	·064	·044	·332	·004	·308	− ·020	·352	·024
34.1	·466 ± ·017	·521 ± ·032	·468	·055	·001	·464	− ·003	·464	− ·002	·466	− ·000
25.1	·473 ± ·017	·425 ± ·036	·466	− ·048	− ·007	·458	− ·015	·469	− ·004	·462	− ·011
35.1	·428 ± ·018	·495 ± ·033	·438	·067	·010	·447	·019	·429	·001	·447	·019
45.1	·681 ± ·012	·759 ± ·019	·715	·078	·034	·751	·069	·747	·066	·686	·005
21.45	·225 ± ·021	·326 ± ·039	·221	·101	− ·004	·279	·053	·261	·036	·244	·018
23.45	·214 ± ·021	·332 ± ·039	·254	·119	·041	·262	·048	·227	·014	·249	·035
13.45	·692 ± ·011	·716 ± ·021	·708	·024	·016	·689	− ·003	·681	− ·011	·700	·008
24.13	·444 ± ·018	·432 ± ·036	·444	− ·012	·000	·446	·002	·449	·005	·440	− ·003
25.13	·390 ± ·019	·289 ± ·040	·363	− ·101	− ·027	·367	− ·023	·392	·002	·364	− ·026
45.13	·602 ± ·014	·676 ± ·024	·642	·073	·040	·685	·083	·685	·082	·603	·001
21.345	·110 ± ·022	·134 ± ·043	·060	·024	·050	·141	·031	·149	·039	·101	− ·009
23.145	·082 ± ·022	·150 ± ·043	·142	·068	·060	·100	·018	·070	− ·012	·113	·031
24.135	·284 ± ·020	·335 ± ·039	·295	·051	·011	·287	− ·003	·269	− ·015	·297	·013
25·134	·171 ± ·021	− ·004 ± ·044	·113	− ·175	− ·058	·094	− ·077	·130	− ·041	·137	− ·034

* The Probable Errors of the partial tetrachorics are approximations obtained by assuming that the probable error of a partial tetrachoric *r* is twice that obtained by the product moment formula. See footnote on p. 255.

Values of β_1 *and* β_2.

	β_1	β_2		β_1	β_2
Vital Capacity	·074 ± ·034	3·375 ± ·230	Weight ...	·6098 ± ·179	4·492 ± ·823
Standing Height	·0006 ± ·0003	3·388 ± ·202	Chest ...	·181 ± ·006	3·452 ± ·234
Sitting Height	·00015 ± ·00004	3·144 ± ·136			

In (*d*) the Biserial *r*'s used were 14, 15, 23, 25, 34 and the Tetrachoric *r*'s were 12, 13, 24, 35, 45.
In (*e*) the Tetrachoric *r*'s used were 12, 13, 24, 35, 45 and the Product Moment *r*'s were 14, 15, 23, 25, 34.
In (*f*) the Biserial *r*'s used were 14, 15, 23, 25, 34 and the Product Moment *r*'s were 12, 13, 24, 35, 45.

If we consider the particular case of this sample, there is only one case —$r_{25.134}$—which might lead to any practical difference in conclusions based on the respective r's and their probable errors*, which is a very favourable result since the number in this sample is large, and with smaller samples deviations of the order observed here would be of still less practical importance compared with the sampling error.

Test II a and *b*. The material used in the second test was taken from "A Study of Index Correlations" by Brown, Greenwood and Wood, published in the *Journal of the Royal Statistical Society*, Vol. LXXVIII. (1913—14). (See Tables II *a* and II *b*.)

TABLE II *a*.

1 = Births, 2 = Population, 3 = Deaths.

Number of Observations = 1000.

Sub-scripts of r	Product Moment	Tetrachoric	Q_5	Tetrachoric – Product Moment	Q_5 – Product Moment
12	·935 ± ·003	·969 ± ·005	·978	·035	·043
13	·733 ± ·010	·938 ± ·008	·941	·204	·208
23	·782 ± ·008	·959 ± ·004	·965	·176	·183
12.3	·852 ± ·006	·713 ± ·021	—	– ·139	—
13.2	·011 ± ·021	·118 ± ·042	—	·108	—
23.1	·400 ± ·018	·583 ± ·028	—	·182	—

Values of β_1 and β_2.

	β_1	β_2
Population	·089	2·023
Births ...	·337	2·665
Deaths ...	16·020	49·819

This material was purposely chosen as being very unsuitable for either tetrachoric or biserial methods. The distributions are artificially restricted at both ends by excluding districts with extreme values of the population, and they are also in some cases extremely skew. In spite of the theoretical inapplicability of

---

* The standard deviation of a partial r derived from coefficients found by any method whatever is given approximately by an equation of the form

$$\sigma^2_{r_{12\cdot3}} \times F_1 = \sigma^2_{r_{12}} \times F_2 + \sigma^2_{r_{13}} \times F_3 + \sigma^2_{r_{23}} \times F_4 + [\delta r_{12}\delta r_{13}]\,F_5 + [\delta r_{12}\delta r_{23}]\,F_6 + [\delta r_{13}\delta r_{23}]\,F_7,$$

where the F's are all functions of the r's (see Dr Heron's note *Biometrika*, Vol. VII. p. 412). If therefore we assume that the mean products $[\delta r_{12}\,\delta r_{13}]$ etc. of the tetrachoric r's are in the same ratio k^2 to the mean products found by the product moment formula (*Phil. Trans.* A, 191, pp. 229 *et seq.*), as the squares of the tetrachoric standard deviations are to the squares of the standard deviations found by the product moment formula, then it follows that the standard deviation of a partial tetrachoric is k times the standard deviation found by the product moment formula. In our data the average value of k for the standard deviations of the total coefficients is 1·85; we have therefore, in order to obtain a rough approximation to the standard deviations of the tetrachoric partials, adopted the above assumption and taken k as 2.

TABLE II *b*.

1 = Birth-Rates, 2 = Population, 3 = Death-Rates.
Number of Observations = 1000.

Sub-scripts of r	Product Moment	Biserial	Biserial − Product Moment
12	·138 ± ·021	·178	·041
13	− ·015 ± ·021	·019	·034
23	·148 ± ·021	·178	·030
12.3	·141 ± ·021	·178	·037
13.2	− ·036 ± ·021	·014	·050
23.1	·151 ± ·021	·177	·026

Values of β_1 *and* β_2.

	β_1	β_2
Birth Rates	·312	5·897
Death Rates	61·539	131·137

the methods, and the fact which is pointed out* by the authors of the paper from which the data were taken, that the material is eminently unsuitable for a basis for practical conclusions, both methods come out of the test well, if we consider the relative errors. Owing to the nature of the distributions, the value of biserial r for population and births came out greater than unity. To test the biserial method we therefore used birth-rates and death-rates instead of births and deaths, these also gave very skew distributions (see Table II *b*).

The average absolute differences here of the tetrachorics were: for the three totals ·138, and for the three partials ·143; these are considerably higher than those in the normal material, but the values of r are here so high that the differences are of no practical importance. The average absolute differences between the biserial and product moment values were: ·035 for the three totals, and ·038 for the three partials; they are large relatively to r but insignificant in this case when the probable errors are taken into account.

Test III. In both Tests I and II*a* the total correlation coefficients found by the product moment method were large.

The data for the third test were taken from 8 diametric measurements of the female pelvis made by Dr Emmons and used by Dr de Souza (*Biometrika*, Vol. IX. 1913, p. 486), since, in contrast to those used in Tests I and II, the product moment coefficients of correlation were fairly small, only 4 out of 28 being larger than ·50.

---

* *loc. cit.* p. 826. The authors' object was to compare the results of two arithmetical processes when the nature of the material was such that a more extreme divergence than was likely to arise in real statistical practice could be anticipated. The paper was *not* intended to be a contribution to the study of the relation between birth-rates and death-rates.

The 28 total coefficients of correlation were found by Dr de Souza for every diameter taken with every other diameter in turn. The total correlation coefficients used here were taken as given in the paper quoted above (see Table III). They are only given there to two places, but it was not thought worth while to recalculate them. In order to avoid introducing further errors, the tetrachorics were found to three places and the final results are shown here to three places, though not accurate to more than two in the case of the product moment coefficients. As before, the dividing lines were taken as near the median as possible, but in the case of some of the variates, e.g. 3 (x_3), 6 (x_6) and 8 (x_8), the division was necessarily very unequal. (The class frequencies in these cases of the 216 observations were: for x_3, 69 and 147; for x_8, 82 and 134; for x_6, 83·5 and 132·5.)

The frequency distributions might all be samples from normal distributions since the number of observations is only 216, but the results for our purposes can be regarded as typical of deviations from distributions whose skewness is described by the observed values of β_1 and β_2 considered apart from their probable errors in this particular case.

The average absolute deviations of the tetrachorics are:

Totals	·065,
Partials, 1st order	·086,
„ 2nd order	·099,
„ 3rd order	·118,

and the largest individual differences of the third order are ·257 and $-$·253. These are, as might be expected, rather larger than those in the more normal material of Test I, and as the third column shows, the relative errors are in some cases very big. In any case, however, where the probable errors are of the order of those in this particular sample, the deviations due to the approximate method would not be larger than those due to random sampling.

In the material of Tests II and III a comparison was also made with Professor Pearson's empirical coefficient Q_5* for a fourfold table:

$$Q_5 = \sin\left(\frac{\pi}{2}\,\frac{1}{\sqrt{1+k^2}}\right) \text{ where } k^2 = \frac{4abcd\,N^2}{(ad-bc)^2\,(a+d)\,(b+c)}.$$

This formula is much quicker to use than the tetrachoric method, and Dr Brownlee kindly lent us his manuscript table of Q_5 for the values of log k^2. The values of Q_5 found in Test III for the 28 total r's hardly differ at all from those of the tetrachorics even when the division is far from median. Even in the very skew material of Test II Q_5 and the tetrachoric r agree much more closely with each other than either does with the product moment, and this suggests that as far as these tests go, the extra time taken in finding the tetrachoric rather than the Q_5 is not compensated for by extra accuracy, especially as in cases where we cannot

---

* *Phil. Trans.* A, 195, 1900.

TABLE III.

Number of Observations = 216.

Subscripts of r	Product Moment	Tetrachoric	Q_5	δr Tetrachoric − Product Moment	$\delta r'$ Q_5 − Product Moment	Tetrachoric $100\frac{\delta r}{r}$
12	·79 ± ·02	·786 ± ·04	·787	− ·004	− ·003	− ·5
13	·61 ± ·03	·533 ± ·06	·565	− ·077	− ·045	− 12·6
14	·24 ± ·04	·207 ± ·07	·209	− ·033	− ·031	− 13·8
15	·17 ± ·04	·115 ± ·07	·115	− ·055	− ·055	− 32·4
16	·30 ± ·04	·134 ± ·07	·136	− ·166	− ·164	− 55·3
17	·10 ± ·05	·053 ± ·07	·053	− ·047	− ·047	− 47·0
18	11 ± ·05	·043 ± ·07	·044	− ·067	− ·066	− 60·9
23	52 ± ·03	·435 ± ·07	·454	·085	− ·066	− 16·4
24	·16 ± ·04	·125 ± ·07	·125	− ·035	− ·035	− 21·9
25	·13 ± ·05	·028 ± ·07	·029	− ·102	− ·101	− 78·5
26	·18 ± ·04	·159 ± ·07	·160	− ·021	− ·020	− 11·7
27	·07 ± ·05	·058 ± ·07	·059	− ·012	− ·011	− 17·1
28	·13 ± ·05	·137 ± ·07	·139	·007	·009	5·4
34	·09 ± ·05	·060 ± ·08	·062	− ·030	− ·028	− 33·3
35	·07 ± ·05	·012 ± ·08	·013	− ·058	− ·057	− 82·9
36	·35 ± ·04	·297 ± ·07	·318	− ·053	− ·032	− 15·1
37	·32 ± ·04	·290 ± ·07	·304	− ·030	− ·016	− 9·4
38	·17 ± ·04	·300 ± ·07	·323	·130	·153	76·5
45	·91 ± ·01	·919 ± ·02	·935	·009	·025	1·0
46	·31 ± ·04	·231 ± ·07	·236	− ·079	− ·074	− 25·5
47	·20 ± ·04	·213 ± ·07	·214	·013	·014	6·5
48	·21 ± ·04	·392 ± ·07	·406	·182	·196	86·7
56	·30 ± ·04	·226 ± ·07	·229	− ·074	− ·071	− 24·7
57	·19 ± ·04	·258 ± ·07	·258	·068	·068	35·8
58	·21 ± ·04	·322 ± ·07	·328	·112	·118	53·3
67	·32 ± ·04	·192 ± ·07	·194	− ·128	− ·126	− 40·0
68	·49 ± ·03	·616 ± ·05	·617	·126	·127	25·7
78	·36 ± ·04	·331 ± ·07	·335	− ·029	− ·025	− 8·1
12.4	·784 ± ·02	·783 ± ·04	—	− ·001	—	− ·2
13.2	·380 ± ·04	·374 ± ·08	—	− ·007	—	− 1·7
13.4	·609 ± ·03	·533 ± ·07	—	− ·076	—	− 12·4
14.2	·188 ± ·05	·193 ± ·09	—	·005	—	2·9
14.6	·162 ± ·05	·183 ± ·09	—	·021	—	12·7
15.4	− ·120 ± ·05	− ·195 ± ·09	—	− ·075	—	62·2
15.6	·088 ± ·05	·088 ± ·09	—	·000	—	0·0
15.3	·161 ± ·05	·128 ± ·09	—	− ·033	—	− 20·3
16.3	·117 ± ·05	− ·030 ± ·09	—	− ·147	—	− 125·8
16.2	·262 ± ·04	·016 ± ·09	—	− ·246	—	− 93·8
17.6	·004 ± ·05	·028 ± ·09	—	·024	—	536·4
18.3	·008 ± ·05	− ·145 ± ·09	—	− ·153	—	− 1887·7
23.4	·514 ± ·03	·432 ± ·08	—	− ·083	—	− 16·1
23.1	·078 ± ·05	·033 ± ·09	—	− ·045	—	− 57·4
25.4	− ·038 ± ·05	− ·222 ± ·09	—	− ·184	—	482·9
26.1	− ·098 ± ·05	·095 ± ·09	—	·193	—	− 197·9
28.1	·071 ± ·05	·182 ± ·09	—	·111	—	157·4
34.2	·008 ± ·05	·006 ± ·09	—	− ·002	—	− 22·2
35.2	·003 ± ·05	− ·0002 ± ·09	—	− ·003	—	− 107·1
35.4	− ·029 ± ·05	− ·110 ± ·09	—	− ·081	—	280·6
36.1	·221 ± ·04	·269 ± ·09	—	·048	—	21·8
36.2	·305 ± ·04	·256 ± ·09	—	− ·049	—	− 16·0
38.1	·131 ± ·05	·328 ± ·08	—	·197	—	150·8
38.2	·121 ± ·05	·270 ± ·09	—	·149	—	122·9
45.6	·901 ± ·01	·915 ± ·02	—	·014	—	1·5
45.3	·910 ± ·01	·920 ± ·01	—	·010	—	1·1
45.2	·909 ± ·01	·923 ± ·01	—	·015	—	1·6
46.3	·299 ± ·04	·224 ± ·09	—	− ·075	—	− 25·1
46.2	·290 ± ·04	·216 ± ·09	—	− ·074	—	− 25·6

TABLE III (*continued*).

Number of Observations = 216.

Sub-scripts of r	Product Moment	Tetrachoric	Q_5	δr Tetrachoric − Product Moment	$\delta r'$ Q_5 − Product Moment	Tetrachoric $\frac{100\,\delta r}{r}$
47.6	·112 ± ·05	·177 ± ·09	—	·065	—	57·8
48.2	·193 ± ·05	·381 ± ·08	—	·188	—	97·3
48.3	·198 ± ·05	·393 ± ·08	—	·194	—	98·0
56.3	·295 ± ·04	·233 ± ·09	—	− ·062	—	− 21·0
57.6	·104 ± ·05	·225 ± ·09	—	·121	—	115·9
58.3	·202 ± ·05	·334 ± ·08	—	·132	—	65·7
58.2	·196 ± ·05	·321 ± ·08	—	·125	—	63·6
68.1	·482 ± ·03	·616 ± ·06	—	·134	—	27·9
68.3	·466 ± ·03	·578 ± ·06	—	·112	—	24·0
12.45	·786 ± ·02	·774 ± ·04	—	− ·012	—	− 1·6
13.45	·610 ± ·03	·525 ± ·07	—	− ·085	—	− 13·9
14.23	·200 ± ·04	·206 ± ·09	—	·006	—	3·0
14.65	·192 ± ·04	·254 ± ·09	—	·062	—	32·5
15.36	·134 ± ·05	·139 ± ·09	—	·006	—	4·3
16.23	·165 ± ·05	− ·089 ± ·09	—	− ·254	—	− 153·8
17.65	− ·005 ± ·05	·009 ± ·09	—	·013	—	− 279·2
18.36	− ·053 ± ·05	− ·156 ± ·09	—	− ·104	—	196·6
23.45	·514 ± ·03	·420 ± ·08	—	− ·094	—	− 18·2
34.25	·013 ± ·05	·017 ± ·09	—	·004	—	28·2
36.12	·230 ± ·04	·267 ± ·09	—	·037	—	16·0
38.12	·126 ± ·05	·327 ± ·08	—	·202	—	160·2
38.25	·123 ± ·05	·285 ± ·08	—	·162	—	132·0
46.23	·302 ± ·04	·221 ± ·09	—	− ·080	—	− 26·6
46.35	·076 ± ·05	·025 ± ·09	—	− ·052	—	− 67·9
47.65	·042 ± ·05	− ·073 ± ·09	—	− ·115	—	− 272·3
48.25	·036 ± ·05	·233 ± ·09	—	·197	—	− 541·9
48.35	·037 ± ·05	·232 ± ·09	—	·195	—	526·8
58.36	·076 ± ·05	·251 ± ·09	—	·175	—	231·4
68.12	·493 ± ·03	·612 ± ·06	—	·120	—	24·3
68.35	·435 ± ·04	·546 ± ·06	—	·111	—	25·6
14.567	·192 ± ·04	·255 ± ·09	—	·063	—	33·0
15.368	·138 ± ·05	·187 ± ·09	—	·049	—	35·2
16.234	·113 ± ·05	− ·141 ± ·09	—	− ·253	—	− 225·2
23.451	·070 ± ·05	·026 ± ·09	—	− ·044	—	− 62·5
38.245	·122 ± ·05	·289 ± ·08	—	·166	—	136·0
48.356	·004 ± ·05	·261 ± ·09	—	·257	—	5969·8
68.123	·480 ± ·05	·576 ± ·06	—	·096	—	20·0

Pelvic Measurements	Values of β_1	Values of β_2
1 = Intercrests	·089 ± ·060	3·01 ± ·269
2 = Interspines	·103 ± ·088	3·34 ± ·451
3 = Transverse	·122 ± ·100	3·34 ± ·499
4 = Diagonal conjugate	·030 ± ·024	3·02 ± ·250
5 = Obstetric conjugate	·043 ± ·035	3·02 ± ·256
6 = Antero-posterior	·051 ± ·040	2·98 ± ·243
7 = Inter-tubers	·038 ± ·029	2·93 ± ·219
8 = Post sagittal	·700 ± ·499	4·90 ± ·284

use the product moment r we are not usually in a position to judge if the material is normal or not.

The whole set of results is summarised in Table IV, grouped according to the size of the product moment r's. There seems to be some tendency in these data for the deviations to be rather larger when the correlation coefficients are smaller.

TABLE IV.

Groupings	Number of Cases					Mean absolute δr for Tetrachoric			Mean absolute $\delta r'$ for Q_5		Mean Tetrachoric $\frac{\delta r \times 100}{r}$
				Q_5		Test I	Test II	Test III	Test II	Test III	
	Test I	Test II	Test III	Test II	Test III						Test III
(*a*) *Total r's*											
·00 to ·29	—	—	16	—	16	—	—	·061	—	·063	36·7
·30 to ·59	8	—	9	—	9	·035	—	·087	—	·078	24·5
·60 to ·99	2	3	3	3	3	·034	·139	·030	·144	·024	4·7
(*b*) *1st Partials*											
·00 to ·29	4	1	28	—	—	·070	·108	·098	—	—	175·5
·30 to ·59	6	1	5	—	—	·049	·182	·077	—	—	17·1
·60 to ·99	2	1	5	—	—	·047	·139	·023	—	—	3·4
(*c*) *2nd Partials*											
·00 to ·29	2	—	15	—	—	·110	—	·106	—	—	176·4
·30 to ·59	2	—	4	—	—	·056	—	·101	—	—	23·7
·60 to ·99	2	—	2	—	—	·050	—	·049	—	—	7·8
(*d*) *3rd Partials*											
·00 to ·29	4	—	6	—	—	·079	—	·139	—	—	1076·9
·30 to ·59	—	—	1	—	—	—	—	·096	—	—	20·0
·60 to ·99	—	—	—	—	—	—	—	—	—	—	—
Totals	32	6	94	3	28						

On the whole the results of these experimental tests are not such as to shake our confidence in partial coefficients derived from fourfold or biserial r's, but, as has been said before, these tests cannot go further than to indicate the average order of the deviations to be expected from material of the degree of skewness investigated, and the practical importance or otherwise of these deviations must be judged by the accuracy required and otherwise obtainable in each individual case.

The following method of attacking the problem theoretically has been suggested by Dr Isserlis. It consists in examining the function representing the relative increment in a partial coefficient corresponding to given small relative increments in the total coefficients to see if this function has any maximum value or values within the possible range, and if so, what such values are.

Consider the algebraic function $r_{12.3} = \dfrac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$; not for any particular population or samples but for all possible values of r_{12}, r_{13}, r_{23} between 0 and ± 1 for which $r_{12.3}$, $r_{23.1}$ and $r_{13.2}$ are also between these limits. Suppose that small increments h_{12}, h_{13}, h_{23} (which for the time being we will consider as arbitrary, examining later our assumptions with reference to the present problem) are given to r_{12}, r_{13} and r_{23} and that $h_{12.3}$ is the corresponding increment in $r_{12.3}$. Then to 1st order terms only:

$$h_{12.3} = \frac{1}{(1-r_{13}^2)^{\frac{1}{2}}(1-r_{23}^2)^{\frac{1}{2}}}\left\{h_{12} - h_{13}\frac{r_{23}-r_{12}r_{13}}{1-r_{13}^2} - h_{23}\frac{r_{13}-r_{12}r_{23}}{1-r_{23}^2}\right\} \quad \ldots\ldots(1),$$

or, taking the relative increment, which is of more practical interest:

$$\frac{h_{12.3}}{r_{12.3}} = \frac{h_{12}}{r_{12}}\,\frac{r_{12}}{r_{12}-r_{13}r_{23}} - \frac{h_{13}}{r_{13}}\,\frac{(r_{23}-r_{13}r_{12})}{(1-r_{13}^2)(r_{12}-r_{13}r_{23})} - \frac{h_{23}}{r_{23}}\,\frac{(r_{13}-r_{12}r_{23})}{(1-r_{23}^2)(r_{12}-r_{13}r_{23})} \quad \ldots(2).$$

If we suppose the relative increments $\frac{h_{12}}{r_{12}}, \frac{h_{13}}{r_{13}}, \frac{h_{23}}{r_{23}}$ to be constant and differentiate expression (2) to find its maxima we reach the result that the only values for which (2) could be a maximum or minimum are either

$$r_{12} = r_{13}r_{23},$$

or $$r_{13} = r_{12} = 0,$$

or $$r_{23} = r_{12} = 0,$$

or $$r_{12} = r_{13} = r_{23} = 0.$$

In each of these cases $r_{12.3} = 0$ so that the relative error is of no interest. The conclusion therefore is that for all the r's within the range 0 to ± 1 the relative error (so far as it is given by the approximate expression (2)) has no maximum value. The same conclusion is found to be true of the absolute error, if in this case we suppose the absolute increments h_{12}, h_{13}, h_{23} to be constant. The function $r_{12.3} = \frac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$ may however be discontinuous for certain values of the r's. When either r_{13} or r_{23} is unity, then r_{12} is equal to either r_{23} or r_{13} as the case may be and the function $\frac{r_{12}-r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$ is indeterminate. Values of either or both of the two secondary correlations r_{13} or r_{23} approaching unity call therefore for fuller consideration. Before going into this case, we will first see how far we are justified in applying the above conclusions to our present problem. The r's are product moment r's and the h's we take to be the differences between the values of r as found by the product moment and—say—the tetrachoric methods; they are not to be confused with either sampling errors, or with the differential increments which we give to the r's in the process of finding a maximum. Our experimental results seem to justify the assumption that the h's are small compared to the r's except when the r's themselves are very small. With regard to our other assumptions, it may be objected that we cannot vary the r's without varying the h's too, and that correlational relations exist between the h's, so that we have no right to assume either $\frac{h_{12}}{r_{12}}$ etc. or h_{12} etc. constant. This would, of course, be perfectly true if our universe or field of variation were confined to samples from one single population and if the h's were sampling errors in either product moment or tetrachoric r's, and it *might* also be true in the same confined universe of samples from a single population if the h's were given their present meaning of differences between tetrachoric and product moment values (we are not able to investigate this theoretically without a knowledge of the frequency distributions of the original variables in the population, and our experimental results do not cover samples from the same population). The point however does not, I think, arise here, since the universe we are considering is not limited to any particular population and its samples, but comprises all possible populations of all degrees of skewness and of correlation likely to arise in practice, with their samples. Hence to any particular set of r values can correspond any number of different values of the h's, these values depending, as we have seen, mainly on the deviations from normality in all the different original distributions rather than on the values of the r's. And *vice versa*, to any particular set of h values can correspond any number of different values of the r's depending on the actual correlation between the variables, which can take all values from 0 to ± 1. Hence it seems not unreasonable to assume that the expression (1) above is a continuous function of the r's in the still wide universe limited only by picking out a definite arbitrary constant set of values of h_{12}, h_{13} and h_{23}, or that (2) is a continuous function of the r's in another universe limited only by picking out definite arbitrary constant values of $\frac{h_{12}}{r_{12}}, \frac{h_{13}}{r_{13}}, \frac{h_{23}}{r_{23}}$.

The fundamental question we ask is therefore—Given certain relative or absolute values of the deviations h, are there any particularly dangerous sets of r to which these can be applied?—dangerous, that is, as regards the size of the resulting relative or absolute deviation in $r_{12.3}$. The answer—only an approximate one—is that so far as regards the value of a quantity which (except near the limits 0 and ± 1 of the total r's) closely approximates to the deviation in question, there are no such dangerous values; hence within these limits we can safely say there are no particularly dangerous ones for the true deviations. We now go back to the cases when either r_{13} or r_{23} approaches unity. Numerical illustrations of the large fluctuations in partial coefficients which can be obtained by giving small *arbitrary* deviations to the total coefficients in these cases are given in Table V. The possibility of such deviations will then be considered.

This table also illustrates the fact, obvious enough *a priori*, that $r_{12.3}$ is not unreliable when r_{12} alone is approximately equal to unity, i.e. to get instability the high total correlation must involve the variable kept constant. The deviations in the total coefficients in Table V are all arbitrary and were chosen to give large deviations in the partials, but they are mostly of about the order of the sampling errors of the respective totals. These considerations led me at first sight into the error of altogether mistrusting partial coefficients derived from such high totals even in the case of product moment r's. That this suspicion of entire unreliability is however not well founded in the case of sampling errors of product moment r's

TABLE V.

Arbitrary Deviations.

		r	δr (arbitrary)	Partials from r	Partials from $r+\delta r$	δr (Partial)	N	χ^2	Approximate order of P
(1)	r_{12}	·96	·03	− ·009	·717	·726	100	28·2	10^{-6}
	r_{13}	·97	·01	·496	− ·184	·680			
	r_{23}	·99	·003	·863	·812	− ·051			
(2)	r_{12}	·40	·05	− ·231	·418	·648	100	31·3	10^{-7}
	r_{13}	·45	− ·05	·318	− ·361	− ·647			
	r_{23}	·98	·01	·997	·990	·013	200	62·7	10^{-13}
(3)	r_{12}	·571600 ± ·083	·0684	− ·473	·112	·585	30	28·5	10^{-6}
	r_{13}	·951313 ± ·012	− ·011313	·943	·897	− ·046			
	r_{23}	·708506 ± ·061	− ·058506	·651	·185	− ·466			
(4)	r_{12}	·809521 ± ·031	− ·029521	·044	− ·347	− ·391	55	6·81	·080
	r_{13}	·964803 ± ·006	·005197	·894	·931	·037			
	r_{23}	·832386 ± ·028	·017614	·333	·614	·281			
(5)	r_{12}	·883242 ± ·027	·026758	− ·630	·508	1·135	30	179	10^{-38}
	r_{13}	·995626 ± ·001	− ·001626	·988	·974	− ·014			
	r_{23}	·911441 ± ·021	− ·021441	·732	− ·321	− 1·053			
(6)	r_{12}	·970 ± ·003	·003	·825	·878	·053	163	29	10^{-6}
	r_{13}	·906 ± ·009	− ·02	·202	− ·276	− ·478			
	r_{23}	·914 ± ·009	·02	·339	·672	·333			

Sources of Data.

(1) and (2) fictitious.

(3) p. 21, (4), p. 22 and (5) p. 29 of E. C. Snow, M.A., "The Intensity of Natural Selection in Man," *Drapers' Co. Research Memoirs, Studies in National Deterioration*, pp. 1—43.

(6) Greenwood and Newbold, *Journal of Hygiene*, Vol. XXI. No. 4, August 1923, Table IV, p. 444.

is seen when the correlation in errors between total coefficients is taken into account, and it was found that though any one of the above deviations in the total r's is not unlikely to arise as an error of sampling, in every case the combination of the three would be a most unlikely event. The last column of Table V gives a rough approximation* to the probability in each case of a set of deviations in the three total r's equally or more improbable than the arbitrary ones given in the second column. These probabilities are negligibly small in every case but one, and even in that case though P is ·080, N is only 55, so that the deviations of $r_{12\cdot 3}$ and $r_{23\cdot 1}$ are not more than 2·88 and 2·32 times their respective standard deviations. These few cases cannot of course prove that the large deviations in the partials would always be found to be so unlikely, but any other result would be inconsistent with the general theorem proved by Yule[*] that the distribution of sampling errors of a partial coefficient is the same as that of a total coefficient. The demonstration of this theorem (which Fisher[†] has since emended by showing that in the case of the partial distribution n should be replaced by $n - s$, where s is the number of variables kept constant) is quite general and should apply just as well to the case where the total coefficients are high as to any other. In fact, this case has been tested experimentally by Bispham[‡] for samples of 30, in which the secondary total correlations were as high as ·977 and the mean partial coefficient approximately zero; he found that his observed distributions of partial coefficients gave a reasonable agreement with the predicted values based on the cooperative paper from the Galton Laboratory[§] on total correlation coefficients.

---

* This approximation has been made by assuming a normal correlation solid in fourfold space for the distribution of errors in the three total coefficients, and finding χ^2 and P in the usual way. The assumption of normality is of course not justified for such high values of r, but in the absence of the requisite knowledge for dealing with the appropriate skew solid, this method may serve as a rough guide. It probably underestimates the value of P, but the values found leave room for this.

TABLE VI.

Observed Deviations due to Tetrachorics.

		r Product Moment	r Tetrachoric	δr Tetrachoric – Product Moment	Partials from Product Moments	Partials from Tetrachorics	δr (Partial)	N	χ^2	Approximate order of P
(1)	r_{12}	·933582 ± ·006	·962250 ± ·012	·028668	·569 ± ·032	·299 ± ·087*	– ·270	200	31·4	10^{-7}
	r_{13}	·900105 ± ·009	·965202 ± ·011	·065097	·807 ± ·017	·739 ± ·043	– ·068			
	r_{23}	·966318 ± ·003	·981356 ± ·007	·015038	– ·022 ± ·048	·399 ± ·080	·421			
(2)	r_{12}	·969810 ± ·003	·953441 ± ·015	– 016369	·825 ± ·017	·665 ± ·059	– ·160	163	90·7	10^{-20}
	r_{13}	·906360 ± ·009	·919303 ± ·022	·012943	·202 ± ·051	– ·225 ± ·100	– ·427			
	r_{23}	·913921 ± ·009	·978574 ± ·008	·064836	·339 ± ·047	·863 ± ·027	·524			

* See footnote on p. 255.

Values of β_1 and β_2.

			β_1	β_2
(1)	Deaths age 0—1 ...	x_1	2·697	5·121
	Deaths age 1—2 ...	x_2	4·224	7·040
	Deaths all ages ...	x_3	6·758	11·197
(2)	Oxygen intake ...	x_1	·0255	2·282
	Carbon dioxide output	x_2	·0103	2·185
	Work per minute ...	x_3	·4078	2·283

For true sampling errors then, there appears to be no danger in high total coefficients. Discussion as to what kinds of deviations can be regarded as true sampling errors is outside the scope of this paper, but it is clear that the deviations arising from the use of tetrachoric instead of product-moment r's are not sampling errors; hence in the case of tetrachoric partials there is no reason why deviations from product moment partials as large as the fictitious ones in Table V should not arise. That they do in fact sometimes arise when high totals are used is shown by the two sets of actual observations given in Table VI.

---

[*] *Proc. Roy. Soc.* A, LXXIX. pp. 182—198, 1907. "On the Theory of Correlation for any number of Variables treated by a new System of Notation," by G. Udny Yule.

[†] *Metron*, Vol. III. No. 3, 1924, pp. 329—333. "The Distribution of the partial Correlation Coefficient," by R. A. Fisher.

[‡] *Metron*, Vol. II. No. 4, 1923, pp. 684—696. "An experimental Determination of the Distribution of the partial Correlation Coefficient in Samples of thirty," by J. W. Bispham.

[§] *Biometrika*, Vol. XI. pp. 328—413. "On the Distribution of the Correlation Coefficient in small Samples." A Cooperative Study by H. E. Soper, A. W. Young, B. M. Cave, A. Lee and K. Pearson.

In the first set all three variables are very skew, in the second set one is moderately skew but none is normal. The first set consists of the male deaths ages 0—1 years, 1—2 years, and all ages respectively in the 10 years 1901—1910 in the first 200 Registration Districts of England as given in the Decennial Supplement, omitting those in which the total male deaths exceeded 9500; the second set consists of the 163 observations of oxygen intake, carbon dioxide output and work done per minute cited above in the last example in Table V. Three out of the six partials have deviations of ·4 to ·5. The deviations from product moment values in the total coefficients clearly do not (and there is no apparent reason why they should) follow the same correlated distributions as sampling errors, since the approximate probabilities of the observed or more unlikely sets arising as errors of sampling are negligibly small (see last column of Table VI). The danger of getting quite unreliable partials from tetrachoric or other approximate methods in cases where the secondary totals are high does therefore exist, but it is not however likely to arise often in practice owing to the comparative rarity of these very high correlations.

Summary.

(1) Experimental tests have been made on three sets of material of varying degrees of abnormality of the deviations from product moment partials, likely to arise in partials derived from total coefficients found by other methods.

(2) The average order of the deviations found is summarised in Table IV.

(3) In the particular samples examined the deviations, though in some cases large, are not of practical importance, compared with the sampling errors.

(4) Search for maximum values of the deviations in partials by the usual algebraic method fails to reveal any particularly dangerous combinations of total coefficients except very high values of secondary totals, and the unreliability of tetrachoric partials in such cases is confirmed by tests on two further sets of observations.

Note on Miss Newbold's Memoir.

The previous paper has been published because it raises a number of most interesting points and has involved clearly a very large amount of computing work. But the Editor considers and he thinks the author would admit that it cannot be looked upon as a final treatment of the topic. It is not only (*a*) that excessively anormal distributions have been used as test material and (*b*) that these are in several cases obtained by very artificial truncations, destroying any approach to linearity, but that the theory which makes h_{12}/r_{12}, h_{13}/r_{13}, h_{23}/r_{23} constants is quite inadequate, even for errors of observation as apart from deviations of sampling. Innumerable quantitative tables each with its own product-moment value of r will give the same tetrachoric r, and a triply infinite series of product-moment partials will give the same tetrachoric partial r. The first steps that are needed in analysis are (i) to determine the sampling distribution of a total tetrachoric r and (ii) to find the correlation between total product-moment r and total tetrachoric r. These problems do not seem beyond the present state of analysis. We can then possibly proceed to the like problems for partial tetrachorics. Meanwhile the Editor ventures to think that a table like Table VI is very liable to misinterpretation. The distributions of product-moment r from ·90 to ·98 etc. are known, and the probable errors provided in the table do not really represent them. Still less is this true for the tetrachoric totals. What may be the distribution of differences between product-moment and tetrachoric r's is a quite unsolved problem.

In practical statistics so much work has been done that we know fairly closely now what is the nature of the distributions for a great variety of characters. We know that deaths with age, births with time, weights with age, population, birth-rates and death-rates do *not* give normal distributions or linear regressions. There is another wide variety of characters which we know do fulfil these conditions. When we come to deal practically with statistical data even when we know the class it is always advisable to study correlation by two (or more, if possible) different methods, and only speak with some confidence of the results if they are confirmatory. It is very rarely that some forms of the correlation ratio, of mean square contingency, of biserial r, etc. cannot be used in conjunction, so that the investigator has some confirmatory evidence of the extent of the association. This remark seems to be called for by the statement on p. 252, that "in a practical case...we have no means of knowing the distribution and may be dealing with a very skew surface."

One last remark, the total tetrachoric coefficient assumes normality in the population sampled. The sample to which we apply the product-moment method is not necessarily normal even if we sample from a normal population. If we take a fourfold sample from a normal population, we can shuffle and *ruffle* the contents of each of the four quadrants in a great variety of ways, so that the normality of the original population is disguised from the standpoint of the product-moment sample, while preserved in the tetrachoric. Such "ruffling," especially in the case of extreme observations due to eccentricities of observation, or errors of nomenclature, or prejudices of the market (in economic matters) or idiosyncrasies of legislators or judges (in criminological data), may render the tetrachoric correlation and its partials, notwithstanding higher probable errors, more reliable than product-moment *r*'s. It is precisely the same conception as that involved in the statement that a median estimate of judgments by a mixed group of jurors is more reliable than their average. This remark has been made because a careless reader of Miss Newbold's paper might conceive that product-moment correlation has some prescriptive pre-eminence, that position can only be given to the correlation ratio and not to the correlation coefficient. Correlation analysis is a name covering an extensive "instrumentarium" and only a wide experience can really enable the statistical worker to determine what tool to use under given circumstances; it will depend not only on the grouping of the data, but on what is often more important the class of data to which they belong and the nature of the recorders. K.P

THE FIFTEEN CONSTANT BIVARIATE FREQUENCY SURFACE.

BY KARL PEARSON.

(1) *Introductory.* At the conclusion of a paper "On a certain Hypergeometrical Series and its Representation by Continuous Frequency Surfaces" (*Biometrika*, Vol. XVI. p. 185), I suggested that we should not get really satisfactory skew-frequency surfaces until we had enough constants at our disposal to determine independently all moments up to the fourth order. The hypergeometrical series with double variation considered in that memoir contained—apart from a constant determined by the volume—only four independent constants, which were thus not enough to give absolute freedom to the correlation r, and the marginal constants β_1, β_2 and β_1', β_2' of the two variates*.

In order to give freedom to r, β_1, β_2, β_1', β_2' it was needful to still further generalise the double hypergeometrical series. This I did in a very simple manner. I supposed a population of N individuals, m of them being marked. I then drew a sample of n individuals and did not return them to the bag. I now placed in the bag m' more marked individuals and drew a sample of n'. If $z_{ss'}$ be the frequency of occurrence of samples of s marked individuals in the first sample of n and s' marked individuals in the second sample of n', then $z_{ss'}$ determined a double hypergeometrical series with five constants N, m, m', n and n' which can be chosen so as to give the r, β_1, β_2, β_1', β_2' of any system of observation.

I failed, however, to obtain integrable differential equations to the surface "parallel" to the spacial histogram thus determined, just as I failed many years ago to integrate the equations resulting from the simpler double hypergeometrical. In despair of achieving any result in this way, I applied—as Laplace had done to the simple single hypergeometrical series—Stirling's Theorem, and evaluated $z_{ss'}$ in a similar manner. The result was a simple normal bivariate surface of which the constants had their usual values in r, σ_1 and σ_2, but the ordinate of this normal surface was multiplied by a polynomial with linear and cubic approximative terms.

I then realised that if I took my skew-frequency surface to be

$$z = z_0 e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)} \left\{1 - a_0 + a_1\frac{x}{\sigma_1} + a_2\frac{y}{\sigma_2} + b_1\frac{x^2}{\sigma_1^2} + 2b_2\frac{xy}{\sigma_1\sigma_2} + b_3\frac{y^2}{\sigma_2^2} + c_1\frac{x^3}{\sigma_1^3} \right.$$
$$\left. + c_2\frac{x^2y}{\sigma_1^2\sigma_2} + c_3\frac{xy^2}{\sigma_1\sigma_2^2} + c_4\frac{y^3}{\sigma_1^3} + d_1\frac{x^4}{\sigma_1^4} + d_2\frac{x^3y}{\sigma_1^3\sigma_2} + 3d_3\frac{x^2y^2}{\sigma_1^2\sigma_2^2} + d_4\frac{xy^3}{\sigma_1\sigma_2^3} + d_5\frac{y^4}{\sigma_2^4}\right\} \quad \text{.........(i)},$$

---

* The position of the double mean and the values of the standard deviations are not here included in the constants, as the first is a question of how the series is imposed on the observations and the latter are really only questions of scale in plotting.

I should have enough constants[*], namely σ_1, σ_2, r, a_0, a_1, a_2, b_1, b_2, b_3, c_1, c_2, c_3, c_4 d_1, d_2, d_3, d_4, d_5, or 18 constants in all to determine:

(i) the total volume; (ii) and (iii) the position of the mean; (iv) and (v) the two variate standard deviations; (vi) the coefficient of correlation; (vii), (viii), (ix) and (x) the four marginal β's, β_1, β_2, β_1', β_2'; (xi) and (xii) the two third order product moment coefficients about the mean; and (xiii), (xiv), (xv) the three fourth order product moment coefficients. This would leave me three constants to spare, and I was therefore at liberty to make the constants I had called r, σ_1 and σ_2 the true correlation and the true standard deviations of my two variates. There would be no difficulty in determining my fifteen constants a_0, a_1, a_2, b_1, b_2, b_3, c_1, c_2, c_3, c_4, d_1, d_2, d_3, d_4 and d_5 for the momental and product momental identities would in every case give linear equations, which would only require straightforward if very laborious algebra for their solution[†].

Such is the history of my attempt to find a bivariate frequency surface which gives all the first fifteen momental constants of any bivariate set of observations. It is, perhaps, needless to observe that I am not over-pleased with the result. I do not lay great stress on the fact that besides the labour of computing the fifteen first momental constants of a set of observations, there will be the great labour of investigating the contour lines. All such labour comes in the day's work of the statistician, and as he certainly cannot express the univariate frequency distribution by less than four momental constants, so he cannot expect to be let off a bivariate distribution with less than fifteen. My main criticism of the surface (i) is that although we may have observations and surface in agreement for the fifteen constants, there is no evidence that this method of expansion provides a series in any way convergent. It certainly must fail for the crateroid and *J*-sectional surfaces, which we have reached in practical statistics. There are many other surfaces also which are theoretically limited by lines or triangles, and which it is impossible to represent by (i) although they would have the same fifteen moments as (i).

But the full investigation of (i) will serve to throw useful light on the legitimacy of expansions of this sort, which have recently been suggested from more than one quarter. This matter we shall discuss as the surface is developed.

(2) *Determination of the Constants of Equation* (i). I shall adopt the following notation. If

$$p_{ss'} = Sn_{xy}\,(x - \bar{x})^s\,(y - \bar{y})^{s'}/N,$$

then I shall write $q_{ss'} = p_{ss'}/(\sigma_1{}^s\sigma_2{}^{s'})$.

Here $\bar{x}$ and $\bar{y}$ are the means, σ_1 and σ_2 the standard deviations of the means of the variates x and y of the total population N, n_{xy} is the frequency of the pair x, y, $= z\,dx\,dy$ when we pass to continuous variation. We shall choose z_0 to be given by

$$z_0 = N/(2\pi\sigma_1\sigma_2\sqrt{1 - r^2})\text{..............................(ii).}$$

---

[*] z_0 is not an independent constant and may be chosen at will.

[†] The algebraic theory of the 15-constant surface with its marginal totals and regression lines was given in the Lent Term of this year (1925) in academic lectures.

For a normal bivariate distribution we have*:

$$q_{01}=q_{10}=0,\quad q_{02}=q_{20}=1,\quad q_{03}=q_{30}=0,\quad q_{04}=q_{40}=3,\quad q_{05}=q_{50}=0,$$
$$q_{06}=q_{60}=15,\quad q_{07}=q_{70}=0,\quad q_{08}=q_{80}=105,\quad q_{11}=r,\quad q_{12}=q_{21}=0,$$
$$q_{13}=q_{31}=3r,\quad q_{14}=q_{41}=0,\quad q_{15}=q_{51}=15r,\quad q_{16}=q_{61}=0,\quad q_{17}=q_{71}=105r,$$
$$q_{22}=1+2r^2,\quad q_{23}=q_{32}=0,\quad q_{24}=q_{42}=3\,(1+4r^2),\quad q_{25}=q_{52}=0,$$
$$q_{26}=q_{62}=15\,(1+6r^2),\quad q_{33}=3r\,(3+2r^2),\quad q_{34}=q_{43}=0,\quad q_{35}=q_{53}=15r\,(3+4r^2),$$
$$q_{44}=3\,(3+24r^2+8r^4)\,\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)}.$$

These are the values that must be used to obtain the integrals, when we integrate the right-hand side of Equation (i). But as we shall never need to use these symbols q on the right-hand side of our results, but always put in their values in terms of the correlation coefficient r, we may reserve the symbols q for the observations themselves, only emphasising that they will not then be restricted to any special values as in the case of normal correlation.

We first note that if we multiply (i) by an odd power or product in x and y, we shall obtain only an equation involving a_1, a_2, c_1, c_2, c_3, c_4. On the other hand, if we multiply by an even power or product in x and y, we shall obtain an equation involving a_0, b_1, b_2, b_3, d_1, d_2, d_3, d_4, d_5.

We next note two further points: (i) that the origin is taken to be the mean, (ii) that in multiplying to integrate we shall multiply by $\left(\frac{x}{\sigma_1}\right)^s\left(\frac{y}{\sigma_2}\right)^{s'} d\left(\frac{x}{\sigma_1}\right) d\left(\frac{y}{\sigma_2}\right)$ (varying s and s'), and always of course integrate between the limits $-\infty$ to $+\infty$ of both variates, although in practice we may have absolute knowledge that x and y can only cover a much more limited range of values†. In such cases we suppose the observed frequency zero without this range, but we still equate the frequency constants for a theoretical surface of infinite spread to the observed constants over a limited spread. We cannot therefore anticipate good results from such a method, when the limits of the observational spread are just the loci where the frequency is most emphasised, e.g. cloudiness at two stations, or vision as tested in the usual way by Snellen's types of right and left eyes‡. There is therefore considerable opening for criticism even in the limits chosen for the integrations.

Multiply (i) by x/σ_1 and integrate, we have by (iii):

$$\int_{-\infty}^{-\infty}\int_{-}^{+\infty} z\frac{x}{\sigma_1}\,d\frac{x}{\sigma_1}\,d\frac{y}{\sigma_1}=N\frac{x}{\sigma_1}=0=N\,\{a_1+ra_2+3c_1+3rc_2+(1+2r^2)\,c_3+3rc_4\}.$$

---

* *Biometrika*, Vol. XII. pp. 86—87.

† For example, correlation between the numbers of cards of the same suit in two players' hands: the possible range is the right-angled triangle of sides 0 to 13 and 0 to 13.

‡ Another good illustration is that of the *percentage* errors made in bisecting a line. Clearly in measuring an individual's accuracy, we can only take a *positive* percentage error, for the individual who makes a negative error is as inaccurate as one who makes a positive error, and we cannot treat the population as a whole as absolutely accurate. Our percentage errors are therefore limited to the range 0 to 50. The maximum frequency is close to zero, or we get a J-shaped distribution. If we now measure the error in trisecting a line the range of percentage error is limited in a singular manner, which I will not enter into here, but there is again a lumping up of frequency in the region of zero and we get for the correlation surface of relation between bisectional and trisectional accuracy in the same individual a J-sectioned frequency surface.

Hence:
$$0 = a_1 + ra_2 + 3c_1 + 3rc_2 + (1 + 2r^2)\,c_3 + 3rc_4 \quad\text{..............(iv).}$$

Similarly for the y mean:
$$0 = ra_1 + a_2 + 3rc_1 + (1 + 2r^2)\,c_2 + 3rc_3 + 3c_4 \quad\text{....... ... .(v).}$$

We now reach in the same manner two equations by multiplying by $(x/\sigma_1)^3$ and by $(y/\sigma_2)^3$, and integrating. We have:
$$\sqrt{\beta_1} = 3a_1 + 3ra_2 + 15c_1 + 15rc_2 + 3\,(1 + 4r^2)\,c_3 + 3r\,(3 + 2r^2)\,c_4 \quad\text{...(vi),}$$
$$\sqrt{\beta_1'} = 3ra_1 + 3a_2 + 3r\,(3 + 2r^2)\,c_1 + 3\,(1 + 4r^2)\,c_2 + 15rc_3 + 15c_4 \quad\text{...(vii).}$$

Here $\sqrt{\beta_1}$ and $\sqrt{\beta_1'}$ must be given the same signs as ${}_x\mu_3$ and ${}_y\mu_3$, the third moment coefficients of the x and y marginal totals.

Lastly we multiply by $\left(\frac{x}{\sigma_1}\right)^2 \frac{y}{\sigma_2}$ and $\frac{x}{\sigma_1}\left(\frac{y}{\sigma_2}\right)^2$ respectively, and thus obtain the values of q_{21} and q_{12} for the observed data. We find:
$$q_{21} = 3ra_1 + (1 + 2r^2)\,a_2 + 15rc_1 + 3\,(1 + 4r^2)\,c_2 + 3r\,(3 + 2r^2)\,c_3 + 3\,(1 + 4r^2)\,c_4 \quad\text{.........(viii),}$$
$$q_{12} = (1 + 2r^2)\,a_1 + 3ra_2 + 3\,(1 + 4r^2)\,c_1 + 3r\,(3 + 2r^2)\,c_2 + 3\,(1 + 4r^2)\,c_3 + 15rc_4 \quad\text{.........(ix).}$$

The six linear equations (iv) to (ix) suffice to determine the six coefficients a_1, a_2, c_1, c_2, c_3, c_4 of the odd powers in the polynomial of Equation (i). We need not detain the reader with the steps in the solution, but place here the results:
$$a_1 = \frac{1}{2\,(1 - r^2)^2}\left\{3rq_{21} - \sqrt{\beta_1} - (1 + 2r^2)\,q_{12} + r\sqrt{\beta_1'}\right\} \quad\text{.................(x),}$$
$$a_2 = \frac{1}{2\,(1 - r^2)^2}\left\{3rq_{12} - \sqrt{\beta_1'} - (1 + 2r^2)\,q_{21} + r\sqrt{\beta_1}\right\} \quad\text{.................(xi),}$$
$$c_1 = \frac{1}{6\,(1 - r^2)^3}\left\{r^2\,(3q_{12} - r\sqrt{\beta_1'}) - (3rq_{21} - \sqrt{\beta_1})\right\} \quad\text{.....................(xii),}$$
$$c_2 = \frac{1}{2\,(1 - r^2)^3}\left\{(1 + 2r^2)\,q_{21} - r\sqrt{\beta_1} - r\,((2 + r^2)\,q_{12} - r\sqrt{\beta_1'})\right\} \quad\text{...(xiii),}$$
$$c_3 = \frac{1}{2\,(1 - r^2)^3}\left\{(1 + 2r^2)\,q_{12} - r\sqrt{\beta_1'} - r\,((2 + r^2)\,q_{21} - r\sqrt{\beta_1})\right\} \quad\text{...(xiv),}$$
$$c_4 = \frac{1}{6\,(1 - r^2)^3}\left\{r^2\,(3q_{21} - r\sqrt{\beta_1}) - (3rq_{12} - \sqrt{\beta_1'})\right\} \quad\text{....................(xv).}$$

For checking numerical work, it may be of service to remark that:
$$a_1 = -3c_1 - 2rc_2 - c_3 \text{ and } a_2 = -c_2 - 2rc_3 - 3c_4 \quad\text{...........(xvi).}$$

We now proceed to the even moments and products. Integrating (i) as it stands, we have:
$$N = N\,(1 - a_0 + b_1 + 2rb_2 + b_3 + 3d_1 + 3rd_2 + 3\,(1 + 2r^2)\,d_3 + 3rd_4 + 3d_5),$$
or
$$a_0 = b_1 + 2rb_2 + b_3 + 3d_1 + 3rd_2 + 3\,(1 + 2r^2)\,d_3 + 3rd_4 + 3d_5 \quad\text{...(xvii).}$$

This gives us a_0 in terms of the eight even constants.

The following equations are obtained by multiplying by x^2/σ_1^2 and y^2/σ_2^2, and integrating:

$$a_0 = 3b_1 + 6rb_2 + (1 + 2r^2)\, b_3 + 15d_1 + 15rd_2 + 9\,(1 + 4r^2)\, d_3 + 3r\,(3 + 2r^2)\, d_4 + 3\,(1 + 4r^2)\, d_5 \quad\text{......(xviii)},$$

$$a_0 = (1 + 2r^2)\, b_1 + 6rb_2 + 3b_3 + 3\,(1 + 4r^2)\, d_1 + 3r\,(3 + 2r^2)\, d_2 + 9\,(1 + 4r^2)\, d_3 + 15rd_4 + 15d_5 \quad\text{......(xix)}.$$

Multiplying by $xy/(\sigma_1\sigma_2)$, we deduce:

$$ra_0 = 3rb_1 + 2\,(1 + 2r^2)\, b_2 + 3rb_3 + 15rd_1 + 3\,(1 + 4r^2)\, d_2 + 9r\,(3 + 2r^2)\, d_3 + 3\,(1 + 4r^2)\, d_4 + 15rd_5 \quad\text{......(xx)}.$$

The fourth moments of the marginal totals provide the equations:

$$\beta_2 - 3 = -3a_0 + 15b_1 + 30rb_2 + 3\,(1 + 4r^2)\, b_3 + 105d_1 + 105rd_2 + 45\,(1 + 6r^2)\, d_3 + 15r\,(3 + 4r^2)\, d_4 + 3\,(3 + 24r^2 + 8r^4)\, d_5 \quad\text{......(xxi)},$$

$$\beta_2' - 3 = -3a_0 + 3\,(1 + 4r^2)\, b_1 + 30rb_2 + 15b_3 + 3\,(3 + 24r^2 + 8r^4)\, d_1 + 15r\,(3 + 4r^2)\, d_2 + 45\,(1 + 6r^2)\, d_3 + 105rd_4 + 105d_5 \quad\text{......(xxii)}.$$

We have now to obtain the fourth order product moments by multiplying by $\dfrac{x^3}{\sigma_1^3}\dfrac{y}{\sigma_1}$, $\dfrac{x^2y^2}{\sigma_1^2\sigma_2^2}$ and $\dfrac{x}{\sigma_1}\dfrac{y^3}{\sigma_2^3}$ respectively, and integrating. We deduce:

$$q_{31} - 3r = -3ra_0 + 15rb_1 + 6\,(1 + 4r^2)\, b_2 + 3r\,(3 + 2r^2)\, b_3 + 105rd_1 + 15\,(1 + 6r^2)\, d_2 + 45r\,(3 + 4r^2)\, d_3 + 3\,(3 + 24r^2 + 8r^4)\, d_4 + 15r\,(3 + 4r^2)\, d_5 \quad\text{......(xxiii)},$$

$$q_{22} - 1 - 2r^2 = -(1 + 2r^2)\, a_0 + 3\,(1 + 4r^2)\, b_1 + 6r\,(3 + 2r^2)\, b_2 + 3\,(1 + 4r^2)\, b_3 + 15\,(1 + 6r^2)\, d_1 + 15r\,(3 + 4r^2)\, d_2 + 9\,(3 + 24r^2 + 8r^4)\, d_3 + 15r\,(3 + 4r^2)\, d_4 + 15\,(1 + 6r^2)\, d_5 \quad\text{......(xxiv)},$$

$$q_{13} - 3r = -3ra_0 + 3r\,(3 + 2r^2)\, b_1 + 6\,(1 + 4r^2)\, b_2 + 15rb_3 + 15r\,(3 + 4r^2)\, d_1 + 3\,(3 + 24r^2 + 8r^4)\, d_2 + 45r\,(3 + 4r^2)\, d_3 + 15\,(1 + 6r^2)\, d_4 + 105rd_5 \quad\text{.........(xxv)}.$$

By aid of (xvii) a_0 was eliminated from Equations (xviii)—(xxv), and these equations were then rewritten with the following abbreviations:

$$\left.\begin{aligned} B_2 = \tfrac{1}{24}\,(\beta_2 - 3), \quad B_2' = \tfrac{1}{24}\,(\beta_2' - 3), \\ Q_{31} = \tfrac{1}{6}\,(q_{31} - 3r), \quad Q_{13} = \tfrac{1}{6}\,(q_{13} - 3r), \quad Q_{22} = \tfrac{1}{12}\,(q_{22} - 1 - 2r^2) \end{aligned}\right\}\text{(xxvi)}.$$

There resulted the following system of equations, of which the first three provided the b's in terms of the d's. These were substituted in the last five, which were then solved for the d's.

$$0 = b_1 + 2rb_2 + r^2b_3 + 6d_1 + 6rd_2 + 3\,(1 + 5r^2)\, d_3 + 3r\,(1 + r^2)\, d_4 + 6r^2d_5 \quad\text{...(xxvii)},$$

$$0 = r^2b_1 + 2rb_2 + b_3 + 6r^2d_1 + 3r\,(1 + r^2)\, d_2 + 3\,(1 + 5r^2)\, d_3 + 6rd_4 + 6d_5 \quad\text{...(xxviii)},$$

$$0 = 2rb_1 + 2\,(1 + r^2)\, b_2 + 2rb_3 + 12rd_1 + 3\,(1 + 3r^2)\, d_2 + 12r\,(2 + r^2)\, d_3 + 3\,(1 + 3r^2)\, d_4 + 12rd_5 \quad\text{...(xxix)},$$

$$2B_2 = b_1 + 2rb_2 + r^2b_3 + 8d_1 + 8rd_2 + 3\,(1 + 7r^2)\, d_3 + r\,(3 + 5r^2)\, d_4 + 2r^2\,(3 + r^2)\, d_5 \quad\text{......(xxx)},$$

$$2B_2' = r^3 b_1 + 2rb_2 + b_3 + 2r^2(3 + r^2)\,d_1 + r(3 + 5r^2)\,d_2 + 3(1 + 7r^2)\,d_3 + 8rd_4 + 8d_5 \qquad \ldots\ldots\text{(xxxi)},$$

$$\begin{aligned} 2Q_{31} = {} & 4rb_1 + 2(1 + 3r^2)\,b_2 + 2r(1 + r^2)\,b_3 + 32rd_1 + (5 + 27r^2)\,d_2 + 6r(7 + 9r^2)\,d_3 \\ & + (3 + 21r^2 + 8r^4)\,d_4 + 4r(3 + 5r^2)\,d_5 \qquad \ldots\ldots\text{(xxxii)}, \end{aligned}$$

$$\begin{aligned} 2Q_{13} = {} & 2r(1 + r^2)\,b_1 + 2(1 + 3r^2)\,b_2 + 4rb_3 + 4r(3 + 5r^2)\,d_1 + (3 + 21r^2 + 8r^4)\,d_2 \\ & + 6r(7 + 9r^2)\,d_3 + (5 + 27r^2)\,d_4 + 32rd_5 \ldots\ldots\text{(xxxiii)}, \end{aligned}$$

$$\begin{aligned} 6Q_{22} = {} & (1 + 5r^2)\,b_1 + r(2 + r^2)\,b_2 + (1 + 5r^2)\,b_3 + 6(1 + 7r^2)\,d_1 + 3r(7 + 9r^2)\,d_2 \\ & + 6(2 + 17r^2 + 5r^4)\,d_3 + 3r(7 + 9r^2)\,d_4 + 6(1 + 7r^2)\,d_5 \ldots\text{(xxxiv)}. \end{aligned}$$

The solutions of Equations (xxvii) to (xxxiv) and then (xvii) are as follows:

$$a_0 = -\frac{3}{(1 - r^2)^2}\{B_2 + B_2' - r(Q_{31} + Q_{13}) + (1 + 2r^2)\,Q_{22}\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xxxv)},$$

$$b_1 = -\frac{3}{(1 - r^2)^3}\{2B_2 + 2r^2B_2' - 2rQ_{31} - r(1 + r^2)\,Q_{13} + (1 + 5r^2)\,Q_{22}\} \qquad \ldots\ldots\ldots\text{(xxxvi)},$$

$$2b_2 = \frac{3}{(1 - r^2)^3}\{4r(B_2 + B_2') - (1 + 3r^2)(Q_{31} + Q_{13}) + 4r(2 + r^2)\,Q_{22}\} \qquad \ldots\ldots\text{(xxxvii)},$$

$$b_3 = -\frac{3}{(1 - r^2)^3}\{2r^2B_2 + 2B_2' - r(1 + r^2)\,Q_{31} - 2rQ_{13} + (1 + 5r^2)\,Q_{22}\} \qquad \ldots\ldots\text{(xxxviii)},$$

$$d_1 = \frac{1}{(1 - r^2)^4}\{B_2 + r^4B_2' - rQ_{31} - r^3Q_{13} + 3r^2Q_{22}\} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xxxix)},$$

$$d_2 = -\frac{1}{(1 - r^2)^4}\{4rB_2 + 4r^3B_2' - (1 + 3r^2)\,Q_{31} - r^2(3 + r^2)\,Q_{13} + 6r(1 + r^2)\,Q_{22}\} \ldots\text{(xl)},$$

$$d_3 = \frac{1}{(1 - r^2)^4}\{2r^2(B_2 + B_2') - r(1 + r^2)(Q_{31} + Q_{13}) + (1 + 4r^2 + r^4)\,Q_{22}\} \qquad \ldots\ldots\text{(xli)},$$

$$d_4 = -\frac{1}{(1 - r^2)^4}\{4r^3B_2 + 4rB_2' - r^2(3 + r^2)\,Q_{31} - (1 + 3r^2)\,Q_{13} + 6r(1 + r^2)\,Q_{22}\} \ldots\text{(xlii)},$$

$$d_5 = \frac{1}{(1 - r^2)^4}\{r^4B_2 + B_2' - r^3Q_{31} - rQ_{13} + 3r^2Q_{22}\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xliii)}.$$

Equations (x) to (xv) and (xxxv) to (xliii) are the complete formal solution of our problem; they enable us, as soon as the constants

$$N,\ \sigma_1,\ \sigma_2,\ \sqrt{\beta_1},\ \sqrt{\beta_1'},\ \beta_2,\ \beta_2',\ r,\ q_{21},\ q_{12},\ q_{31},\ q_{13},\ q_{22}$$

and the means $\bar{x}$ and $\bar{y}$ are known for our data, to write down a surface, with the whole of these fifteen constants—i.e. the complete set of constants up to moments of the fourth degree identically the same as those of our given data.

(3) *Transformation of the Equation* (i) *into a Second Form.* If $\tilde{z}$ be the function:

$$\tilde{z} = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1 - r^2}}\,e^{-\frac{1}{2}\frac{1}{1 - r^2}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)} \ldots\ldots\ldots\ldots\ldots\text{(xliv)},$$

and we write $x' = x/\sigma_1$, $y' = y/\sigma_2$, it is perfectly easy to write down the various differential coefficients:

$$\frac{d^{s+s'}\tilde{z}}{dx'^s\,dy'^{s'}}.$$

These differential coefficients are, or at least are proportional to, the generalised tetrachoric functions for two correlated variables. Now if we form all the differential coefficients up to the fourth order, it is easy to arrange our surface in the form

$$z = \tilde{z} - \frac{1}{6}\left(\sqrt{\beta_1}\,\frac{d^3\tilde{z}}{dx'^3} + 3q_{21}\,\frac{d^3\tilde{z}}{dx'^2dy'} + 3q_{12}\,\frac{d^3\tilde{z}}{dx'\,dy'^2} + \sqrt{\beta_1'}\,\frac{d^3\tilde{z}}{dy'^3}\right) + B_2\,\frac{d^4\tilde{z}}{dx'^4} + Q_{31}\,\frac{d^4\tilde{z}}{dx'^3dy'}$$

$$+ 3Q_{22}\,\frac{d^4\tilde{z}}{dx'^2dy'^2} + Q_{13}\,\frac{d^4\tilde{z}}{dx'\,dy'^3} + B_2'\,\frac{d^4\tilde{z}}{dy'^4}\,\ldots\ldots\ldots\text{(xlv)}.$$

We now begin to realise some of the defects of the proposed surface. Why should we not introduce the six fifth order bivariate tetrachoric functions? These would enable us to state a surface which would have all the 21 momental coefficients up to and including the fifth order coefficients, and the same process might be continued *ad infinitum.*

The only defence for not doing so would lie in a demonstration that the cubic terms are sensibly greater than the fourth order terms. In the data with which we have practically to deal, the terms B_2, Q_{31}, $3Q_{22}$, Q_{13} and B_2' are not only not very much smaller than $\frac{1}{6}\sqrt{\beta_1}$, $\frac{1}{2}q_{21}$, $\frac{1}{2}q_{12}$ and $\frac{1}{6}\sqrt{\beta_1'}$, but are generally of the same order, and often, in the case of approximately symmetrical surfaces, of a much higher order. If there be no convergency in the series, the idea that the cubic terms will suffice fails completely; it is equivalent to assuming that fourth order terms are essentially normal (i.e. $\beta_2 = 3$, $\beta_2' = 3$, $q_{31} = q_{13} = 3r$, and $q_{22} = 1 + 2r^2$), but the third order terms are not. Actual frequency distributions diverge just as much in their fourth order terms from normality as in their third order terms, i.e. we are no more likely to have $\beta_2 = \beta_2' = 3$ or $q_{22} = 1 + 2r^2$, than $\sqrt{\beta_1} = \sqrt{\beta_1'} = 0$ or $q_{21} = 0$.

Another important criticism will be obvious on examination of our Equations (x) to (xv) and (xxxv) to (xliii); it will be seen that the constants of the polynomial proceed by increasing powers of the inverse of $(1 - r^2)$. Hence there is an actual source of divergence in the series, especially potent when r is large. We have no right to drop out a_0 and retain a_1 and a_2, to drop out the quadratic terms b_1, $2b_2$ and b_3 and retain c_1, c_2, c_3 and c_4. This source of divergence is screened when the polynomial is replaced by the differential functions*, which only shorten the length of the expression but do not assist computation.

It may be said that precisely the same criticisms will apply to the omission of fifth and higher differential functions. This may readily be admitted at least by one who has no *a priori* affection for these expansions in terms of functions derived from the normal surface. But there is a certain betterment for the following reasons:

(i) We cannot possibly get decent fits to univariate frequency distributions without going as far as the fourth moment. Hence we are unlikely to get decent fits to frequency surfaces unless we go as far as the fourth momental coefficients.

(ii) In all these cases theoretical relations hold between the higher momental coefficients and the momental coefficients selected for fitting. But the higher momental coefficients have larger and larger probable errors. Hence the further up we can push the series for selected moments the more likely it is that the theoretical relations referred to will be satisfied closely enough for practical work.

---

* There has been no attempt yet to tabulate the bivariate tetrachoric functions, yet without tabulation they are of small service. But if it ever be attempted the influence of the $1/(1 - r^2)$ factor will soon make itself felt.

The real test is: Can we get anything like goodness of fit from an identity of 15 momental constants? I am convinced that taking only 6 momental constants (i.e. neglecting a_0, b_1, b_2, b_3, d_1, d_2, d_3, d_4, d_5) leads to an *impasse*, for it does not allow for any deviation in kurtosis from normality.

(4) *Marginal Totals.* I proceed now to discuss certain properties, starting with the marginal totals. I integrated (i) for values of y from $-\infty$ to $+\infty$, and so reached the curve of the x-totals. It was:

$$\begin{aligned} z_{x\cdot} = \frac{N}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} \Big\{ & 1 - a_0 + b_3 (1 - r^2) + 3 (1 - r^2)^2 d_5 \\ & + \frac{x}{\sigma_1} (a_1 + r a_2 + c_3 (1 - r^2) + 3r (1 - r^2) c_4) \\ & + \frac{x^2}{\sigma_1^2} (b_1 + 2r b_2 + r^2 b_3 + 3 (1 - r^2)(d_3 + r d_4) + 6r^2 (1 - r^2) d_5) \\ & + \frac{x^3}{\sigma_1^3} (c_1 + r c_2 + r^2 c_3 + r^3 c_4) \\ & + \frac{x^4}{\sigma_1^4} (d_1 + r d_2 + 3r^2 d_3 + r^3 d_4 + r^4 d_5) \Big\}. \end{aligned}$$

Substituting the values of the constants, r disappears, and this reduces to the anticipated form:

$$\begin{aligned} z_{x\cdot} = \frac{N}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} \Big\{ 1 + \tfrac{1}{8} (\beta_2 - 3) - \tfrac{1}{2} \sqrt{\beta_1} \frac{x}{\sigma_1} - \tfrac{1}{4} (\beta_2 - 3) \frac{x^2}{\sigma_1^2} & \\ + \tfrac{1}{6} \sqrt{\beta_1} \frac{x^3}{\sigma_1^3} + \tfrac{1}{24} (\beta_2 - 3) \frac{x^4}{\sigma_1^4} \Big\} & \end{aligned}$$

which may be written:

$$z_{x\cdot} = \frac{N}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} \left\{ 1 + \tfrac{1}{6} \sqrt{\beta_1} \left(\frac{x^3}{\sigma_1^3} - 3 \frac{x}{\sigma_1} \right) + \tfrac{1}{24} (\beta_2 - 3) \left(\frac{x^4}{\sigma_1^4} - 6 \frac{x^2}{\sigma_1^2} + 3 \right) \right\} \quad \text{.........(xlvi).}$$

The corresponding y-marginal curve is

$$z_{\cdot y} = \frac{N}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}} \left\{ 1 + \tfrac{1}{6} \sqrt{\beta_1'} \left(\frac{y^3}{\sigma_2^3} - 3 \frac{y}{\sigma_2} \right) + \tfrac{1}{24} (\beta_2' - 3) \left(\frac{y^4}{\sigma_2^4} - 6 \frac{y^2}{\sigma_2^2} + 3 \right) \right\} \quad \text{.........(xlvii).}$$

The factors $\frac{1}{6}\sqrt{\beta_1}$, $\frac{1}{24}(\beta_2 - 3)$, $\frac{1}{6}\sqrt{\beta_1'}$, $\frac{1}{24}(\beta_2' - 3)$ are seen to be multiplied by the simple fourth and fifth order tetrachoric functions in x and y.

If we define the sth order tetrachoric function to be

$$\left. \begin{aligned} \tau_s(x) &= \frac{1}{\sqrt{s!}} \left(-\frac{d}{dx} \right)^{s-1} \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{s!}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} p_{s-1}(x) \end{aligned} \right\} \quad \text{...............(xlviii),}$$

where

$$p_{s-1}(x) = \left\{x^{s-1} - \frac{(s-1)(s-2)}{2\,.\,1!}\,x^{s-3} + \frac{(s-1)(s-2)(s-3)(s-4)}{2^2\,.\,2!}\,x^{s-5} - \text{etc.}\right\} \qquad \text{.........(xlix)},$$

it is easy to expand any expression in tetrachoric functions, if its momental coefficients be known. For these functions are semi-normal functions, i.e.

$$\int_{-\infty}^{+\infty} \tau_s(x)\,\tau_{s'}(x)\,e^{\frac{1}{2}x^2}\,dx = 0,$$

if s be not equal to s', and

$$\int_{-\infty}^{+\infty} \tau_s^2(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2\pi}}\,\frac{1}{s} \qquad \text{..........................(l)}.$$

But no satisfactory proof of the convergence of such expansions has yet been given. If $F(x)$ be any function of x:

$$\int_{-\infty}^{+\infty} F(x)\,\tau_s(x)\,e^{\frac{1}{2}x^2}\,dx = \int_{-\infty}^{+\infty} F(x)\,p_{s-1}(x)\,dx \times \frac{1}{\sqrt{s!}\,\sqrt{2\pi}}$$

$$= \frac{1}{\sqrt{s!}}\,\frac{1}{\sqrt{2\pi}}\left\{\mu'_{s-1} - \frac{(s-1)(s-2)}{2\,.\,1!}\,\mu'_{s-3} + \frac{(s-1)(s-2)(s-3)(s-4)}{2^2\,.\,2!}\,\mu'_{s-5} - \text{etc.}\right\} \qquad \text{.........(li)}.$$

Hence for successive s's we have

$$\int_{-\infty}^{+\infty} F(x)\,\tau_1(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2\pi}}\,\mu_0' = \frac{1}{\sqrt{2\pi}},$$

$$\int_{-\infty}^{+\infty} F(x)\,\tau_2(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{2}}\,\frac{1}{\sqrt{2\pi}}\,\mu_1' = 0,$$

$$\int_{-\infty}^{+\infty} F(x)\,\tau_3(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{6}}\,\frac{1}{\sqrt{2\pi}}\,(\mu_2' - \mu_0') = 0,$$

$$\int_{-\infty}^{+\infty} F(x)\,\tau_4(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{24}}\,\frac{1}{\sqrt{2\pi}}\,(\mu_3' - 3\mu_1') = \frac{1}{\sqrt{24}}\,\frac{1}{\sqrt{2\pi}}\,\sqrt{\beta_1},$$

$$\int_{-\infty}^{+\infty} F(x)\,\tau_5(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{120}}\,\frac{1}{\sqrt{2\pi}}\,(\mu_4' - 6\mu_2' + 3\mu_0') = \frac{1}{\sqrt{120}}\,\frac{\beta_2 - 3}{\sqrt{2\pi}},$$

$$\int_{-\infty}^{+\infty} F(x)\,\tau_6(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{720}}\,\frac{1}{\sqrt{2\pi}}\,(\mu_5' - 10\mu_3' + 15\mu_1') = \frac{1}{\sqrt{720}}\,\frac{{}^{*}\beta_3' - 10\sqrt{\beta_1}}{\sqrt{2\pi}},$$

$$\int_{-\infty}^{+\infty} F(x)\,\tau_7(x)\,e^{\frac{1}{2}x^2}\,dx = \frac{1}{\sqrt{5040}}\,\frac{1}{\sqrt{2\pi}}\,(\mu_6' - 15\mu_4' + 45\mu_2' - 15\mu_0')$$

$$- \frac{1}{\sqrt{5040}}\,\frac{1}{\sqrt{2\pi}}\,(\beta_4 - 15\beta_2 + 30)$$

$$= \frac{1}{\sqrt{5040}}\,\frac{1}{\sqrt{2\pi}}\,(\beta_4 - 15 - 15\,(\beta_2 - 3)) \qquad \text{...........(lii)},$$

and so on.

---

* $\beta_3' = \mu_5' = \mu_5/\sigma_1^5$ here and is not the more usual $\beta_3 = \mu_5'\mu_3' = \mu_5\mu_3/\sigma_1^8$.

We have accordingly the following result, when we take $z_x = F(x)$, any function of x:

$$\frac{z_x}{N/\sigma_1} = \tau_1 + \frac{4}{\sqrt{24}}\sqrt{\beta_1}\,\tau_4 + \frac{5}{\sqrt{120}}(\beta_2 - 3)\,\tau_5 + \frac{6}{\sqrt{720}}(\beta_3' - 10\sqrt{\beta_1})\,\tau_6 + \frac{7}{\sqrt{5040}}(\beta_4 - 15 - 15(\beta_2 - 3))\,\tau_7 + \text{etc.}\ldots\ldots\ldots(\text{liii}).$$

In this case x/σ_1 will be the argument of the tetrachoric functions, and we must take

$$\tau_s\left(\frac{x}{\sigma_1}\right) = \frac{1}{\sqrt{s!}}\left(-\frac{d}{d\left(\frac{x}{\sigma_1}\right)}\right)^{s-1}\frac{1}{\sqrt{2\pi}\,\sigma_1}\,e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}.$$

With my definition of the tetrachoric function, which introduces the factor $\frac{1}{\sqrt{s!}}$ and so differs from continental usage, all these functions, τ_1 excluded, are of essentially the *same* order*. We cannot therefore assert convergence in these functions in themselves. Next consider the numerical terms, i.e. $s/\sqrt{s!}$; we have for 4, 5, 6, 7, etc. the values:

$$\cdot 816{,}497,\quad \cdot 456{,}435,\quad \cdot 223{,}607,\quad \cdot 098{,}601,\quad \text{etc.}$$

The second of these terms is 55 $^\circ/_\circ$ of the first, and the fourth of them more than 12 $^\circ/_\circ$. There are few cases in which we can afford to neglect 12 $^\circ/_\circ$; none in which we can neglect 55 $^\circ/_\circ$! In fact the sum of the numerical coefficients following the τ_4 term amounts to 100 $^\circ/_\circ$ of that of the coefficient of that term. On the basis of the numerical factors only we cannot assert that it is justifiable to retain τ_4 and neglect τ_5, τ_6, τ_7, etc.

Thus our belief in the convergency of this tetrachoric expansion must depend upon terms like

$$\sqrt{\beta_1},\quad (\beta_2 - 3),\quad (\beta_3' - 10\sqrt{\beta_1}),\quad (\beta_4 - 15 - 15(\beta_2 - 3)),\quad \text{etc.}$$

forming a rapidly converging series.

Now these constants are all *linear* measures of the deviation of the function from a normal curve of errors. Now it will be found that in a very large number of practical cases these normal curve excesses, so far from converging, tend to diverge. I will illustrate this by a simple case. The curve

$$y = y_0\left(1 + \cdot 351{,}364\,\frac{x}{\sigma}\right)^9 e^{-8\cdot 162{,}278\frac{x}{\sigma}}$$

falls quite within the bounds of ordinary everyday statistical experience. It gives

$$\beta_1 = \cdot 4,\quad \beta_2 = 3\cdot 6,\quad \beta_3' = 7\cdot 08350 \text{ and } \beta_4 = 29\cdot 20.$$

Hence the normal curve excess factors are respectively:

$$\cdot 632{,}456,\quad \cdot 600{,}000,\quad \cdot 758{,}942 \text{ and } 5\cdot 200{,}000.$$

---

* Compare the numerical values of these functions in *Tables for Statisticians and Biometricians*, Table XXIX.

These show obvious signs of divergence, which are emphasised if we go to the terms involving τ_8 and τ_9. The source of this is the simple fact, that if the low β's diverge at all from the normal, it will be found that the higher β's diverge in a far higher proportion. Experience shows no convergency in the normal curve excess factors.

If we multiply the above values by the corresponding $s/\sqrt{s!}$ factors, we obtain the series:

$$\tau_1 + \cdot 516{,}398\tau_4 + \cdot 273{,}861\tau_5 + \cdot 169{,}705\tau_6 + \cdot 512{,}725\tau_7 + \text{etc.}$$

Here it is clear that the contribution of τ_7 may be as important as that of τ_4, and that it is quite impossible to retain the τ_4 term and neglect the τ_5, τ_6 and τ_7 terms. There is no convergency about these tetrachoric expansions. It is therefore unlikely that our surface, which is really an expansion in bivariate tetrachorics, is going to give highly satisfactory results, notwithstanding its fifteen momental identities.

It is needful for the application of tetrachoric expansions that the higher normal excess factors should be shown to grow less and less. This in practical experience they certainly do not. How then has arisen the idea that such expansions are reasonable? I think it has arisen from forming hypergeometrical series for urn-problems (as for Bayes' Theorem in the manner of Laplace), and then approximating to the terms of such series by Stirling's Theorem, with such limiting hypotheses as that the sample was small compared to the size of the population in the urn, etc., etc. Good illustrations may be found in the manner in which Edgeworth deduces his skew curve of frequency[*], which amounts to

$$z_x = \tau_1 + a_4\tau_4,$$

and a somewhat similar proof is given by Bowley[†]. In both cases, it is assumed that there are n elemental groups, or what I should term "contributory cause-groups," and these are supposed uncorrelated. Subject to the relation that μ_p/σ^p is finite for all values of p in the contributory cause-groups,

$$z_x = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} \left(1 + \frac{\sqrt{\bar{\beta}_1}}{6\sqrt{n}} \sqrt{24}\,\tau_4 + \frac{\bar{\beta}_2 - 3}{24n} \sqrt{120}\,\tau_5 + \text{etc.}\right),$$

where $\bar{\beta}_1$ and $\bar{\beta}_2$ are the mean β-constants for the contributory cause-groups. It is then argued that the coefficient of τ_5 is of the order $\frac{1}{\sqrt{n}}$ as compared with the coefficient of τ_4, and may therefore be neglected.

Now in the first place the assumption that μ_p/σ^p less its normal value tends to a small finite value is not legitimate. For fairly frequent values of β_1 and β_2, i.e. 0·5 and 4, we have $\beta_5 = 120$, and $\beta_6 = 878$, when for normal values they would be zero and 105. In other words if low β's diverge somewhat from the normal, high values diverge with great rapidity, far more rapidly in fact than terms in $1/\sqrt{n}$, unless n is very large, descend. But the main assumption that frequency distributions are in general due to an indefinitely large number of small contributory cause-groups is not only unproven but can be demonstrated to be incorrect. Professors Edgeworth and Bowley make the factors multiplying τ_4 and τ_5 to contain the mean values of μ_3/σ^3 and $\mu_4/\sigma^4 - 3$ for the contributory groups. But why should the contributory groups have their skewness all in the same sense? Why should all the contributory causes be mesokurtic or leptokurtic? If they are not the ratio of the mean of values of μ_3/σ^3 to that of values of $\mu_4/\sigma^4 - 3$ may take any value whatever, and there is no proof at all that:

---

[*] "The Law of Error." *Cambridge Philosophical Transactions*, Vol xx. pp. 36—65, 1904.

[†] *Elements of Statistics*, p. 291, 1920.

$$\frac{1}{\sqrt{n}}\left(\frac{\bar{\mu}_3}{\sigma^3}\right) \text{ is indefinitely larger than } \frac{1}{n}\left(\frac{\mu_4}{\sigma^4}\right),$$

where the rules stand for mean value. This argument becomes still more cogent when we bear in mind that, except by way of hypothesis, we have no real evidence of this indefinitely large multiplicity of contributory cause-groups. It was called in question by Galton many years ago*. And we can put to a definite proof its legitimacy. The ratio of the coefficient of τ_5 to τ_4, according to Edgeworth and Bowley, is:

$$\frac{\sqrt{5}}{4}\frac{1}{\sqrt{n}}\frac{\bar{\beta}_2 - 3}{\sqrt{\bar{\beta}_1}}, \text{ but } \frac{\sqrt{5}}{4} = \text{actually } \frac{\beta_2 - 3}{\sqrt{\beta_1}},$$

by our analysis above. Hence, if n be indefinitely large, $\beta_2 - 3$ must be indefinitely small as compared with $\sqrt{\beta_1}$; but this is a result contrary to all experience. $\beta_2 - 3$ may have any ratio as compared to $\sqrt{\beta_1}$, and if we had on the basis of observed data to make any statement, all we could hazard would be that these quantities are very much of the same order. Thus either the hypothesis of an indefinitely large number of contributory cause-groups is false, or else the mean β_1 for those groups is indefinitely small, i.e. $\sqrt{\bar{\beta}_1}$ is negligible for those groups as compared to $\bar{\beta}_2 - 3$. In either case the hypothesis adopted by Edgeworth and Bowley is contradicted. The tetrachoric τ_5 term is *not* negligible as compared with τ_4.

It is, of course, equally legitimate to argue that the τ_6 and higher terms cannot be neglected. We can only take comfort by hoping that a curve which has the first four moments identical with those of the observed data will be to some extent better than one which has only three.

It may, of course, be retorted that the skew curves introduced by the present writer only introduce four moment coefficients, and therefore there is no *a priori* reason for supposing them to be in a position to give better results than the curves based on tetrachoric expansions. The answer is fairly straightforward. While the tetrachoric expansions do not converge in terms of excess or defect from the normal distribution, the skew curves in question do give convergence in terms of these excesses or defects.

---

* "Statistics by Intercomparison with Remarks on the Law of Frequency of Error." *Phil. Mag.* Vol. XLIX. pp. 39, 40, 1875.

We start from the differential equation referred to its mean, namely:

$$\frac{1}{y}\frac{dy}{dx/\sigma} = -\frac{\frac{x}{\sigma}+a}{b_0 + b_1\frac{x}{\sigma} + b_2\frac{x^2}{\sigma^2} + b_3\frac{x^3}{\sigma^3} + \dots}.$$

Here b_0 is of the order unity, a and b_1 of the order $\sqrt{\beta_1} = \mu_3/\sigma^3$, and b_2 of the order β_1 and $\beta_2 - 3$. Let us write $\sqrt{\beta_1} = \beta_1'$, and $\beta_3' = \mu_5/\sigma^5$, $\beta_4 = \mu_6/\sigma^6$, and β_2 as usual μ_4/σ^4. Then if we proceed to find our series of b's up to b_3 in the usual manner*, we obtain

$$b_3 = \frac{28\beta_2\beta_1' + 4\beta_2^2\beta_1' + 2\beta_2\beta_3' - 18\beta_3' - 6\beta_3'\beta_1'^2 - 12\beta_1'^3}{\left\{\begin{array}{c}70\beta_4\beta_2 - 126\beta_4 - 100\beta_2^3 + 210\beta_2^2 - 54\beta_2 - 84\beta_4\beta_1'^2 + 60\beta_2'^2\beta_1'^2 - 490\beta_2\beta_1'^2 \\ + 141\beta_3'\beta_1'\beta_2 - 18\beta_1'^2 - 72\beta_3'^2 + 297\beta_1'\beta_3' + 192\beta_1'^4\end{array}\right\}} \quad \text{.........(liv)}.$$

Here the denominator approaches 2 × 144 as the curve approaches the normal, i.e. is finite and large. The numerator can be rearranged as:

$$2\left\{26(\beta_2-3)\beta_1' + (\beta_2-3)\beta_3' - 6\beta_1'\left(\frac{\beta_3'}{\beta_1'} - 10\right) \right.$$
$$\left. + 2(\beta_2-3)^2\beta_1' - 3\beta_3'\beta_1^2 - 6\beta_1^3\right\}.$$

Here all the terms in the second line are of the cubic order; in the first line, the first two are obviously of the square order. We shall now show that the third term is also of the square order. Consider the binomial $(p+q)^n$, we have:

$$\mu_2 = npq = \sigma^2, \quad \mu_3 = npq(p-q) \text{ and } \mu_5 = npq(1 + 2pq(5n-6)).$$

Hence: $$\beta_1' = \frac{npq(p-q)}{(npq)^{\frac{3}{2}}}, \quad \beta_3' = \frac{npq(p-q)(1+2pq(5n-6))}{(npq)^{\frac{5}{2}}}.$$

Accordingly: $$\frac{\beta_3'}{\beta_1'} = \frac{1}{npq} + 10 - \frac{12}{n},$$

or, when n is made indefinitely large:

$$\beta_3'/\beta_1' = 10.$$

But, if n be made indefinitely large, and neither p nor q be very small, the binomial passes over into the normal curve. Accordingly the ratio β_3'/β_1' approaches 10 under the condition of approach to the normal curve, and $\beta_1'(\beta_3'/\beta_1' - 10)$ is therefore of the second order of small quantities. This convergency of the terms† of the series b_0, b_1, b_2, b_3, etc. explains why the types of curves obtained from the above differential equation give far more satisfactory results than tetrachoric expansions such as those proposed by Edgeworth, Charlier and others, which are actually divergent.

---

* "On the General Theory of Skew Correlation and Non-Linear Regression," p. 5. *Drapers' Company Research Memoirs*, 1905, Cambridge University Press.

† The statement of the convergency of this series is not novel, it was made in 1905. I have had many years in my possession the curves resulting from retaining higher denominator powers than the second, but except for a few extreme cases the changes are too slight to be of any profit in dealing with actual frequency distributions.

(5) *Form of the Regression Curves.* Clearly if $\tilde{y}_x$ be the mean of the x-array of y's:

$$n_x \tilde{y}_x = dx \int_{-\infty}^{+\infty} zy\,dy.$$

The integration was performed as before by writing $y - r\dfrac{\sigma_2}{\sigma_1}x = Y$, and integrating with regard to Y. The integration was somewhat laborious, but after substitution of the constants and making certain transformations, I found:

$$\frac{\tilde{y}_x}{\sigma_2} = r\frac{x}{\sigma_1} + \frac{\frac{1}{2}(q_{21} - r\sqrt{\beta_1})\left(\dfrac{x^2}{\sigma_1^2} - 1\right) + \frac{1}{6}(q_{31} - r\beta_2)\left(\dfrac{x^3}{\sigma_1^3} - 3\dfrac{x}{\sigma_1}\right)}{1 + \frac{1}{6}\sqrt{\beta_1}\left(\dfrac{x^3}{\sigma_1^3} - 3\dfrac{x}{\sigma_1}\right) + \frac{1}{24}(\beta_2 - 3)\left(\dfrac{x^4}{\sigma_1^4} - 6\dfrac{x^2}{\sigma_1^2} + 3\right)} \quad \ldots(\text{lv}).$$

Assuming the distribution differs little from the normal the denominator under the long rule may be put unity, and the regression curve will run:

$$\frac{\tilde{y}_x}{\sigma_2} = r\frac{x}{\sigma_1} + \tfrac{1}{2}(q_{21} - r\sqrt{\beta_1})\left(\frac{x^2}{\sigma_1^2} - 1\right) + \tfrac{1}{6}(q_{31} - r\beta_2)\left(\frac{x^3}{\sigma_1^3} - 3\frac{x}{\sigma_1}\right) \ldots(\text{lv})\ \textit{bis}.$$

If $q_{31} - r\beta_2 = 0$ we have parabolic regression, but as a rule there is no more reason for neglecting p_3 than p_2*; indeed experience shows that a cubic is generally more the type of curve needful to describe regression than a parabola. If the surface be not close to the normal we cannot neglect the p_3 and p_4 terms in the denominator, and the regression curve will accordingly then be a quintic. There is not much difficulty, however, in computing $\tilde{y}_x$ for a given x; for multiplying the numerator and denominator by $\dfrac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}$, we can read the result:

$$\frac{\tilde{y}_x}{\sigma_2} = r\frac{x}{\sigma_1} + \frac{\sqrt{\frac{3}{2}}(q_{21} - r\sqrt{\beta_1})\tau_3 + \sqrt{\frac{2}{3}}(q_{31} - r\beta_2)\tau_4}{\tau_1 + \sqrt{\frac{2}{3}}\sqrt{\beta_1}\tau_4 + \sqrt{\frac{5}{24}}(\beta_2 - 3)\tau_5} \quad \ldots\ldots\ldots\ldots(\text{lvi})$$

The coefficients of the tetrachoric functions will be provided as numerics by the data and the values of τ_1, τ_3, τ_4 and τ_5 can be taken at once from the tables. It is accordingly not hard to test the regression curve. It should be noted that when $x = 0$,

$$\tilde{y}_x = -\tfrac{1}{2}\sigma_2(q_{21} - r\sqrt{\beta_1})/(1 + \tfrac{1}{8}(\beta_2 - 3)),$$

or the regression curve does not pass through the mean (i.e. the origin $y = 0$).

Exception may be taken to (lvi) on precisely the same grounds as to the Equations (xlv) and (xlvi), namely, that the stopping at τ_4 and τ_5 is not based on any principle of convergence.

Equation (lv) provides the customary conditions for linearity of regression, i.e.

$$q_{21} = r\sqrt{\beta_1}, \quad q_{31} = r\beta_2.$$

There will be corresponding values for the regression of x on y, if that also is linear, i.e.

$$q_{12} = r\sqrt{\beta_1'}, \quad q_{13} = r\beta_2'.$$

---

* See Equation (xlix) for value of p_{s-1}.

(6) *On the Scedastic Curve for the Arrays.* We need to find the integral

$$\int_{-\infty}^{+\infty} z\,(y-\tilde{y}_x)^2\,dy = z_x \sigma_{y_x}^2$$

$$= \int_{-\infty}^{+\infty} z \left(y - r\frac{\sigma_2}{\sigma_1}x - \left(\tilde{y}_x - r\frac{\sigma_2}{\sigma_1}x\right)\right)^2 dy$$

$$= \int_{-\infty}^{+\infty} z\,(Y^2 - \bar{Y}_x^2)\,dY = \int_{-\infty}^{+\infty} zY^2 dY - z_x \bar{Y}_x^2.$$

The $\int_{-\infty}^{+\infty} zY^2 dY$ was obtained as before by substituting $Y + r\frac{\sigma_2}{\sigma_1}x$ for y in z and integrating out.

After substituting the constant values, I found

$$\int_{-\infty}^{+\infty} zY^2 dY = \frac{N\sigma_2^2}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}$$

$$\times \left\{(1-r^2)\left[1 + \tfrac{1}{6}\sqrt{\beta_1}\left(\frac{x^3}{\sigma_1^3} - 3\frac{x}{\sigma_1}\right) + \tfrac{1}{24}(\beta_2 - 3)\left(\frac{x^4}{\sigma_1^4} - 6\frac{x^2}{\sigma_1^2} + 3\right)\right]\right.$$

$$- \frac{x}{\sigma_1}\{r(q_{21} - r\sqrt{\beta_1}) - (q_{12} - rq_{21})\}$$

$$\left. - \left(\frac{x^2}{\sigma_1^2} - 1\right)\left(r(q_{31} - r\beta_2) - \tfrac{1}{2}\{q_{22} - 1 - r^2(\beta_2 - 1)\}\right)\right\}.$$

Hence dividing by z_x from (xlvi):

$$\frac{1}{z_x}\int_{-\infty}^{+\infty} zY^2 dY = \sigma_2^3(1 - r^2)$$

$$- \sigma_2^2 \frac{\frac{x}{\sigma_1}\{r(q_{21} - r\sqrt{\beta_1}) - (q_{12} - rq_{21})\} + \left(\frac{x^2}{\sigma_1^2} - 1\right)\left(r(q_{31} - r\beta_2) - \tfrac{1}{2}\{q_{22} - 1 - r^2(\beta_2 - 1)\}\right)}{1 + \tfrac{1}{6}\sqrt{\beta_1}\left(\frac{x^3}{\sigma_1^3} - 3\frac{x}{\sigma_1}\right) + \tfrac{1}{24}(\beta_2 - 3)\left(\frac{x^4}{\sigma_1^4} - 6\frac{x^2}{\sigma_1^2} + 3\right)}.$$

To obtain $\sigma_{y_x}^2$ we must subtract from this result $\left(\tilde{y}_x - r\frac{\sigma_2}{\sigma_1}x\right)^2$, which is given by (liv).

Thus we have finally:

$$\sigma_{y_x}^2 = \sigma_2^2(1 - r^2)$$

$$- \sigma_2^2 \frac{\frac{x}{\sigma_1}\{r(q_{21} - r\sqrt{\beta_1}) - (q_{12} - rq_{21})\} + \left(\frac{x^2}{\sigma_1^2} - 1\right)\left(r(q_{31} - r\beta_2) - \tfrac{1}{2}\{q_{22} - 1 - r^2(\beta_2 - 1)\}\right)}{1 + \tfrac{1}{6}\sqrt{\beta_1}\left(\frac{x^3}{\sigma_1^3} - 3\frac{x}{\sigma_1}\right) + \tfrac{1}{24}(\beta_2 - 3)\left(\frac{x^4}{\sigma_1^4} - 6\frac{x^2}{\sigma_1^2} + 3\right)}$$

$$\left. - \frac{\left(\frac{x^2}{\sigma_1^2} - 1\right)\tfrac{1}{2}(q_{21} - r\sqrt{\beta_1}) + \tfrac{1}{6}(q_{31} - r\beta_2)\left(\frac{x^3}{\sigma_1^3} - 3\frac{x}{\sigma_1}\right)}{1 + \tfrac{1}{6}\sqrt{\beta_1}\left(\frac{x^3}{\sigma_1^3} - 3\frac{x}{\sigma_1}\right) + \tfrac{1}{24}(\beta_2 - 3)\left(\frac{x^4}{\sigma_1^4} - 6\frac{x^2}{\sigma_1^2} + 3\right)}\right]^2 \quad \text{(lvii)}.$$

This gives a curve of the ninth order for the variance and one of the tenth order for the standard deviation of the arrays. The computation is not so serious as might at first appear, as $\tilde{y}_x - r\frac{\sigma_2}{\sigma_1}x$ will really have been computed in determining the regression line, and if multiplied by $\frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}$ we can again express our results in tetrachoric functions, which are tabled. Thus

$$\sigma_{y_x}^2 = \sigma_2^2 \Bigg\{ 1 - r^2 - \frac{\sqrt{2}\,\{r(q_{21} - r\sqrt{\beta_1}) - (q_{12} - rq_{21})\}\,\tau_2 + \sqrt{3}\left(r(q_{31} - r\beta_2) - \frac{1}{2}\{q_{22} - 1 - r^2(\beta_2 - 1)\}\right)\tau_3}{\tau_1 + \sqrt{\frac{2}{3}}\sqrt{\beta_1}\,\tau_4 + \sqrt{\frac{5}{24}}\,(\beta_2 - 3)\,\tau_5} - \left[\frac{\sqrt{\frac{3}{2}}\,(q_{21} - r\sqrt{\beta_1})\,\tau_3 + \sqrt{\frac{2}{3}}\,(q_{31} - r\beta_2)\,\tau_4}{\tau_1 + \sqrt{\frac{2}{3}}\sqrt{\beta_1}\,\tau_4 + \sqrt{\frac{5}{24}}\,(\beta_2 - 3)\,\tau_5}\right]^2 \Bigg\} \quad \text{.................................(lviii).}$$

If the regression be linear, $q_{21} = r\sqrt{\beta_1}$, $q_{31} = r\beta_2$, and:

$$\sigma_{y_x}^2 = \sigma_2^2 \left\{ 1 - r^2 + \frac{\sqrt{2}\,(\sqrt{\beta_1'} - r\sqrt{\beta_1})\,r\tau_2 + \frac{1}{2}\sqrt{3}\,\{q_{22} - 1 - r^2(\beta_2 - 1)\}\,\tau_3}{\tau_1 + \sqrt{\frac{2}{3}}\sqrt{\beta_1}\,\tau_4 + \sqrt{\frac{5}{24}}\,(\beta_2 - 3)\,\tau_5} \right\} \quad \text{.........(lix),}$$

or if the deviation from normality be not too great:

$$\sigma_{y_x}^2 = \sigma_2^2 \left\{ 1 - r^2 + \frac{rx}{\sigma_1}(\sqrt{\beta_1'} - r\sqrt{\beta_1}) + \frac{1}{2}\left(\frac{x^2}{\sigma_1^2} - 1\right)\{q_{22} - 1 - r^2(\beta_2 - 1)\} \right\} \quad \text{.........(lx),}$$

which is parabolic scedasticity for the variance, elliptic or hyperbolic for the standard deviation, according as q_{22} is in defect or excess of its normal value.

(7) *Special Cases of some Importance.* Let us assume the regression to be linear; we will afterwards suppose the surface symmetrical as well. The latter case is that of a generalised whist surface.

If the regression be linear:

$$q_{21} = r\sqrt{\beta_1}, \quad q_{12} = r\sqrt{\beta_1'}, \quad q_{31} = \beta_2 r, \quad q_{13} = \beta_2' r \quad \text{............(lxi).}$$

Accordingly:

$$B_2 = \tfrac{1}{24}(\beta_2 - 3), \quad B_2' = \tfrac{1}{24}(\beta_2' - 3), \quad Q_{31} = \tfrac{1}{6}r(\beta_2 - 3), \quad Q_{13} = \tfrac{1}{6}r(\beta_2' - 3),$$
$$Q_{22} = \tfrac{1}{12}(q_{22} - 1 - 2r^2) \quad \text{......(lxii).}$$

We now turn to Equations (x) to (xv). We find:

For Linear Regression.

$$a_1 = \frac{1}{2(1 - r^2)^2}\{\sqrt{\beta_1}(3r^2 - 1) - 2\sqrt{\beta_1'}\,r^3\},$$
$$a_2 = \frac{1}{2(1 - r^2)^2}\{\sqrt{\beta_1'}(3r^2 - 1) - 2\sqrt{\beta_1}\,r^3\},$$
$$c_1 = \frac{1}{6(1 - r^2)^3}\{\sqrt{\beta_1'}\,2r^3 - \sqrt{\beta_1}(3r^2 - 1)\},$$
$$c_2 = \frac{r^2}{2(1 - r^2)^3}\{2r\sqrt{\beta_1} - (1 + r^2)\sqrt{\beta_1'}\},$$
$$c_3 = \frac{r^2}{2(1 - r^2)^3}\{2r\sqrt{\beta_1'} - (1 + r^2)\sqrt{\beta_1}\},$$
$$c_4 = \frac{1}{6(1 - r^2)^3}\{\sqrt{\beta_1}\,2r^3 - \sqrt{\beta_1'}(3r^2 - 1)\},$$
$$\text{.........(lxiii).}$$

For Linear Regression and Sym

$$a_1 = -\frac{1 + 2r}{2(1 + r)^2}\sqrt{\beta_1},$$
$$a_2 = -\frac{1 + 2r}{2(1 + r)^2}\sqrt{\beta_1},$$
$$c_1 = \frac{(1 + 2r)\sqrt{\beta_1}}{6(1 - r^2)(1 + r)^2},$$
$$c_2 = -\frac{r^2\sqrt{\beta_1}}{2(1 - r^2)(1 + r)^2},$$
$$c_3 = -\frac{r^2\sqrt{\beta_1}}{2(1 - r^2)(1 + r)^2},$$
$$c_4 = \frac{(1 + 2r)\sqrt{\beta_1}}{6(1 - r^2)(1 + r)^2},$$
$$\text{.........(lxiv).}$$

We next turn to Equations (xxxv) to (xliii). We deduce, writing $\beta_2 - 3 = \epsilon$, $\beta_2' - 3 = \epsilon'$, $q_{22} - 1 - 2r^2 = \zeta$:

For Linear Regression.

$$\begin{aligned}
a_0 &= -\frac{1}{8(1-r^2)^2}\left\{(\epsilon+\epsilon')(1-4r^2) + 2\zeta(1+2r^2)\right\},\\
b_1 &= -\frac{1}{4(1-r^2)^3}\left\{(1-4r^2)\epsilon - r^2(1+2r^2)\epsilon' + (1+5r^2)\zeta\right\},\\
2b_2 &= \frac{1}{2(1-r^2)^3}\left\{-3r^3(\epsilon+\epsilon') + 2r(2+r^2)\zeta\right\},\\
b_3 &= -\frac{1}{4(1-r^2)^3}\left\{-r^2(1+2r^2)\epsilon + (1-4r^2)\epsilon' + (1+5r^2)\zeta\right\},\\
d_1 &= \frac{1}{24(1-r^2)^4}\left\{(1-4r^2)\epsilon - 3r^4\epsilon' + 6r^2\zeta\right\},\\
d_2 &= -\frac{1}{6(1-r^2)^4}\left\{-3r^3\epsilon - r^3(2+r^2)\epsilon' + 3r(1+r^2)\zeta\right\},\\
d_3 &= \frac{1}{12(1-r^2)^4}\left\{-r^2(1+2r^2)(\epsilon+\epsilon') + (1+4r^2+r^4)\zeta\right\},\\
d_4 &= -\frac{1}{6(1-r^2)^4}\left\{-r^3(2+r^2)\epsilon - 3r^3\epsilon' + 3r(1+r^2)\zeta\right\},\\
d_5 &= \frac{1}{24(1-r^2)^4}\left\{-3r^4\epsilon + (1-4r^2)\epsilon' + 6r^2\zeta\right\},
\end{aligned} \qquad \text{.........(lxv).}$$

For Linear Regression and Symmetry.

$$\begin{aligned}
a_0 &= -\frac{1}{4(1-r^2)^2}\left\{\epsilon(1-4r^2) + \zeta(1+2r^2)\right\},\\
b_1 &= -\frac{1}{4(1-r^2)^3}\left\{(1-5r^2-2r^4)\epsilon + (1+5r^2)\zeta\right\},\\
2b_2 &= \frac{1}{(1-r^2)^3}\left\{-3r^3\epsilon + r(2+r^2)\zeta\right\},\\
b_3 &= -\frac{1}{4(1-r^2)^3}\left\{(1-5r^2-2r^4)\epsilon + (1+5r^2)\zeta\right\},\\
d_1 &= \frac{1}{24(1-r^2)^4}\left\{(1-4r^2-3r^4)\epsilon + 6r^2\zeta\right\},\\
d_2 &= -\frac{1}{6(1-r^2)^4}\left\{-r^3(5+r^2)\epsilon + 3r(1+r^2)\zeta\right\},\\
d_3 &= \frac{1}{12(1-r^2)^4}\left\{-2r^2(1+2r^2)\epsilon + (1+4r^2+r^4)\zeta\right\},\\
d_4 &= -\frac{1}{6(1-r^2)^4}\left\{-r^3(5+r^2)\epsilon + 3r(1+r^2)\zeta\right\},\\
d_5 &= \frac{1}{24(1-r^2)^4}\left\{(1-4r^2-3r^4)\epsilon + 6r^2\zeta\right\},
\end{aligned} \qquad \text{.........(lxvi).}$$

It will be clear that in the first case all we need are mean, standard deviation, and the two β's for both marginal totals, besides r and q_{22}. For the second we need only the constants for one marginal total, and r and q_{22}.

I propose to illustrate now these results on one or two examples in order to test the value of such a 15-constant frequency surface.

(8) *Illustration I. Symmetrical Surface with Linear Regression.* As an illustration, I take the correlation surface for the numbers of the same suit in two hands at whist*. Here we have:

$$\beta_1 = \beta_1' = \cdot 036{,}2667, \quad \beta_2 = \beta_2' = 2\cdot 893{,}2914, \quad r = -\tfrac{1}{3}, \quad \sigma_1^2 = \sigma_2^2 = 1\cdot 863{,}9706,$$
$$\bar{x} = \bar{y} = 3\cdot 25, \quad \sigma_1 = \sigma_2 = 1\cdot 365{,}2731, \quad \sqrt{\beta_1} = \sqrt{\beta_1'} = \cdot 1904{,}3818,$$
$$\zeta = q_{22} - 1 - 2r^2 = -\cdot 044{,}0423, \quad \epsilon = \epsilon' = -\cdot 106{,}7086.$$

There is little doubt that q_{22} could be ascertained from an analysis of the double hypergeometrical; but to save wasting time over lengthy algebra, I computed q_{22} from the Table for *theoretical* whist frequencies on p. 186 of *Biometrika*, Vol. XVI. It may be of interest to put on record the value of q_{22} as deduced from q_{22}' for any arbitrary origin, when the surface has linear regression and is symmetrical. We have

$$q_{22} = q_{22}' - 4r\sqrt{\beta_1}\,\frac{\bar{x}}{\sigma_1} - 2(1+2r)\frac{\bar{x}^2}{\sigma_1^2} - \frac{\bar{x}^4}{\sigma_1^4} \text{ ....(lxvii).}$$

---

* See *Biometrika*, Vol. XVI. p. 176.

To obtain this, it was needful to show that

$$q_{21}' = q_{12}' = r\sqrt{\beta_1} + (1 + 2r)\frac{\bar{x}}{\sigma_1} + \frac{\bar{x}^3}{\sigma_1^3} \quad\ldots\ldots\ldots\ldots\ldots\ldots(\text{lxviii}).$$

In both cases we supposed the arbitrary origin to be at a symmetrical point, i.e. $\bar{x} = \bar{y}$. In the actual computations this was taken at 3, 3, or $\bar{x}$ and $\bar{y} = {\cdot}25$.

The total frequency being 25,000, we have the following numerical value for the equation to the 15-constant surface:

$$\begin{aligned} z = 2264{\cdot}109\, e^{-\frac{9}{16}\left(\frac{x^2}{\sigma_1^2} + \frac{2xy}{3\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)} \Bigg\{ & {\cdot}964{,}2106 - {\cdot}071{,}4143\left(\frac{x}{\sigma_1} + \frac{y}{\sigma_2}\right) \\ & + {\cdot}040{,}3305\left(\frac{x^2}{\sigma_1^2} + {\cdot}675{,}5851\frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right) \\ & + {\cdot}026{,}7804\left(\frac{x^3}{\sigma_1^3} - \frac{x^2y}{\sigma_1^2\sigma_2} - \frac{xy^2}{\sigma_1\sigma_2^2} + \frac{y^3}{\sigma_2^3}\right) \\ - \Bigg(& {\cdot}005{,}6525\frac{x^4}{\sigma_1^4} + {\cdot}007{,}6716\frac{x^3y}{\sigma_1^3\sigma_2} + {\cdot}014{,}0870\frac{x^2y^2}{\sigma_1^2\sigma_2^2} + {\cdot}007{,}6716\frac{xy^3}{\sigma_1\sigma_2^3} \\ & + {\cdot}005{,}6525\frac{y^4}{\sigma_2^4}\Bigg)\Bigg\} \quad\ldots(\text{lxix}). \end{aligned}$$

We now proceeded to find the contour lines of this surface.

The accompanying table exhibits: (A.) the frequency of appearance of certain numbers of cards of the same suit in two players' hands calculated from the double hypergeometrical series of theory, (S.O.) the ordinates of the 15-constant surface, and (S.V.) the corresponding volumes of that surface deduced by the formulae provided in the Appendix to this paper. It will be seen that the ordinates agree much better with the results of the series (A.) than the volumes. Nor is this to be wondered at, for the moments were naturally calculated from discrete quantities (i.e. actual numbers of cards) and not corrected for grouping. We now proceeded to plot the sections of the 15-constant surface (S.O.) precisely as a year ago we plotted the contour lines of the double hypergeometrical series and the two surfaces to the same data*. The results were astonishingly smooth and closely resembled the contours for (A.). In fact the two families of curves were so close that they could not be represented as in the previous cases as superposed. Such goodness of fit had not been anticipated.

The whist contours with cancrine forms differ widely from the ellipses of the normal surface. It was therefore most encouraging, considering the non-normal character of the whist contour, to see how nearly the 15-constant surface bent its contour system to shapes distinctly non-normal. It was also satisfactory to learn that the *non-converging* double-tetrachoric expansion could so satisfactorily fit data provided we equated at least fifteen-moment constants. Of course the negative frequencies show themselves and these would be more conspicuous had we carried

---

* *Biometrika*, Vol. XVI. pp. 178—181.

TABLE I.

Comparison of Ordinates and Volumes of 15-*Constant Surface with Frequencies of Double Hypergeometrical Series for Whist Returns.*

Number of Cards in First Player's Hand of Suit X.

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
0 A. S. O. S. V.	0 ·28 ·65	5 4·41 6·79	24 25·00 32·41	60 66·53 79·22	88 94·38 108·00	79 81·85$^+$ 91·55	44 47·32 51·92	16 17·25$^+$ 18·93	3 3·55$^-$ 4·09	0 ·38 ·51	0 ·02 ·03	0 ·00 ·00	0 ·00 ·00	0 ·00 ·00
1 A. S. O. S. V.	5 4·41 6·79	51 45·91 55·94	212 204·79 223·54	457 456·22 471·99	572 563·11 569·34	433 419·94 422·89	202 197·62 201·22	58 55·79 59·13	10 8·77 10·20	1 ·73 1·01	0 ·03 ·06	0 ·00 ·00	0 ·00 ·00	
2 A. S. O. S. V.	24 25·00 32·41	212 204·79 223·54	748 744·87 752·46	1372 1381·61 1350·38	1444 1436·86 1393·96	910 903·12 884·83	347 348·38 351·72	79 78·52 84·68	10 9·70 11·89	1 ·63 ·96	0 ·02 ·04	0 ·00 ·00		
3 A. S. O. S. V.	60 66·53 79·22	457 456·22 471·99	1372 1381·61 1350·38	2118 2140·00 2054·15	1853 1856·85$^+$ 1791·98	953 960·05$^+$ 949·64	289 297·31 309·30	50 52·62 60·24	5 5·01 6·81	0 ·24 ·44	0 ·01 ·02			
4 A. S. O. S. V.	88 94·38 108·00	572 563·11 569·34	1444 1436·86 1393·96	1853 1856·85$^+$ 1791·98	1324 1316·86 1296·81	542 538·09 554·90	126 126·46 141·86	16 16·63 21·51	1 1·13 1·88	0 ·04 ·05				
5 A. S. O. S. V.	79 81·85$^+$ 91·55	433 419·94 422·89	910 903·12 884·83	953 960·05$^+$ 949·64	542 538·09 554·90	170 163·39 183·15	28 26·51 34·53	2 2·17 3·70	0 ·06 ·21					
6 A. S. O. S. V.	44 47·32 51·92	202 197·62 201·22	347 348·38 351·72	289 297·31 309·30	126 126·46 141·86	28 26·51 34·53	3 2·29 4·24	0 −·01 ·21						
7 A S. O. S. V.	16 17·25$^+$ 18·93	58 55·79 59·13	79 78·52 84·68	50 52·62 60·24	16 16·63 21·51	2 2·17 3·70	0 −·01 ·21							
8 A. S. O. S. V.	3 3·55$^-$ 4·09	10 8·77 10·20	10 9·70 11·89	5 5·01 6·81	1 1·13 1·88	0 ·06 ·21								
9 A. S. O. S. V.	0 ·38 ·51	1 ·73 1·01	1 ·63 ·96	0 ·24 ·44	0 ·04 ·05									
10 A. S. O. S. V.	0 ·02 ·03	0 ·03 ·06	0 ·02 ·04	0 ·01 ·02										
11, 12, 13 A., S. O., S. V.	0 ·00 ·00	0 ·00 ·00	0 ·00 ·00											

our computation of ordinates somewhat further, especially to negative card frequencies.

But this first illustration is undoubtedly valuable as showing that at any rate in frequency surfaces with a single axis of symmetry the 15-constant surface will provide quite a reasonable graduation.

I next proceeded to consider the marginal totals. Now it must be remembered that the moments and product moments of the surface were obtained by treating the frequencies as discrete quantities (number of cards of given suit in hand), and accordingly the frequency surface is what has been termed a "spurious surface" and will really represent by its ordinates rather than by its volumes the cell contents (see p. 285).

The curve of marginal totals is

$$z_x = \frac{N}{\sigma_1}\left\{\tau_1 + \frac{4}{\sqrt{24}}\sqrt{\beta_1}\,\tau_4 + \frac{5}{\sqrt{120}}(\beta_2 - 3)\,\tau_5 + \dots\right\}.$$

Here:

$$\frac{1}{\sigma_1} = \cdot 7324{,}5419, \quad \sqrt{\beta_1} = \cdot 1904{,}3818, \quad (\beta_2 - 3) = -\cdot 106{,}7086,$$

or,
$$z_x = 18311\cdot 355\,(\tau_1 + \cdot 1554{,}9212\,\tau_4 - \cdot 0487{,}0559\,\tau_5) \quad \dots\dots\dots\dots\dots\text{(lxx)},$$

from which Column C of Table II has been calculated.

It will be seen that the sum of the ordinates of the surface in each array is very close to the corresponding marginal ordinate. The fit as measured by P is very good. At the same time it must be remembered that we are not dealing with material showing random deviations due to sampling, but with a mathematical function. The P's as found for the marginal totals in the previous investigation of this surface* are:

$P = \cdot 000$, putting $\beta_1 = 0$, and giving the true value of β_2,

$P = \cdot 350$, putting $\beta_1 =$ the true value, but neglecting β_2.

There is thus very considerable improvement obtained by using the tetrachoric expansion and both β's correct.

If we use a Type I Pearson-Curve which corresponds to the true values of the β's we find for its equation:

$$y = 7181\cdot 642\left(1 - \frac{x}{11\cdot 00731}\right)^{21\cdot 18554}\left(1 + \frac{x}{5\cdot 96003}\right)^{11\cdot 44407} \quad \dots\dots\text{(lxxi)},$$

the origin being at the mode 3·10404 cards.

---

* I have recomputed the P's, using the *ordinates* and a total of 25,000 deals not the 1000 of the earlier investigation, *Biometrika*, Vol. XVI. p. 185.

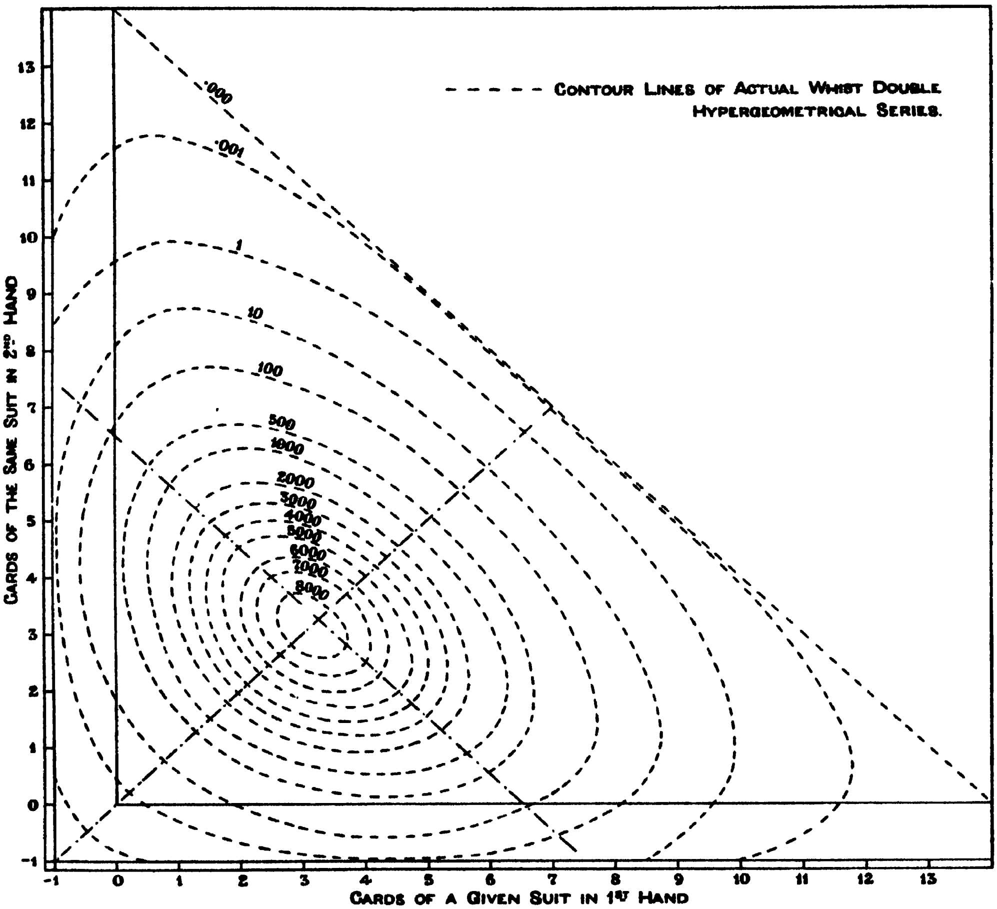


Diagram I.

We have the following results:

TABLE II.

Number of Cards of given suit	A Theoretical Frequencies (i.e. terms of Hypergeometrical)	B Sum of Ordinates of Surface in Array	C Theoretical Curve	Δ = C − A	Contribution to χ^2
0	319·8	340·97	341·0	+21·2	1·405
1	2001·6	1957·32	1956·4	−45·2	1·021
2	5146·8	5133·50	5134·1	−12·7	·031
3	7158·2	7216·45	7218·1	+59·9	·501
4	5965·2	5954·41	5952·7	−12·5	·026
5	3117·3	3095·18	3095·4	−21·9	154
6	1039·1	1045·88	1046·6	+ 7·5	·054
7	220·4	222·97	223·7	+ 3·3	·049
8	29·2	28·22	28·6	− 0·6	·012
9	2·3	2·02	2·1	− 0·2	·017
10	0·1	·08	0·1	·00	·000
11	·0002	·00	·00	·00	—
12	·0000	·00	·00	·00	—
13	·0000	·00	·00	·00	—

χ^2 (for $n'=11$) $=3{\cdot}270$, $P=\cdot 972$.

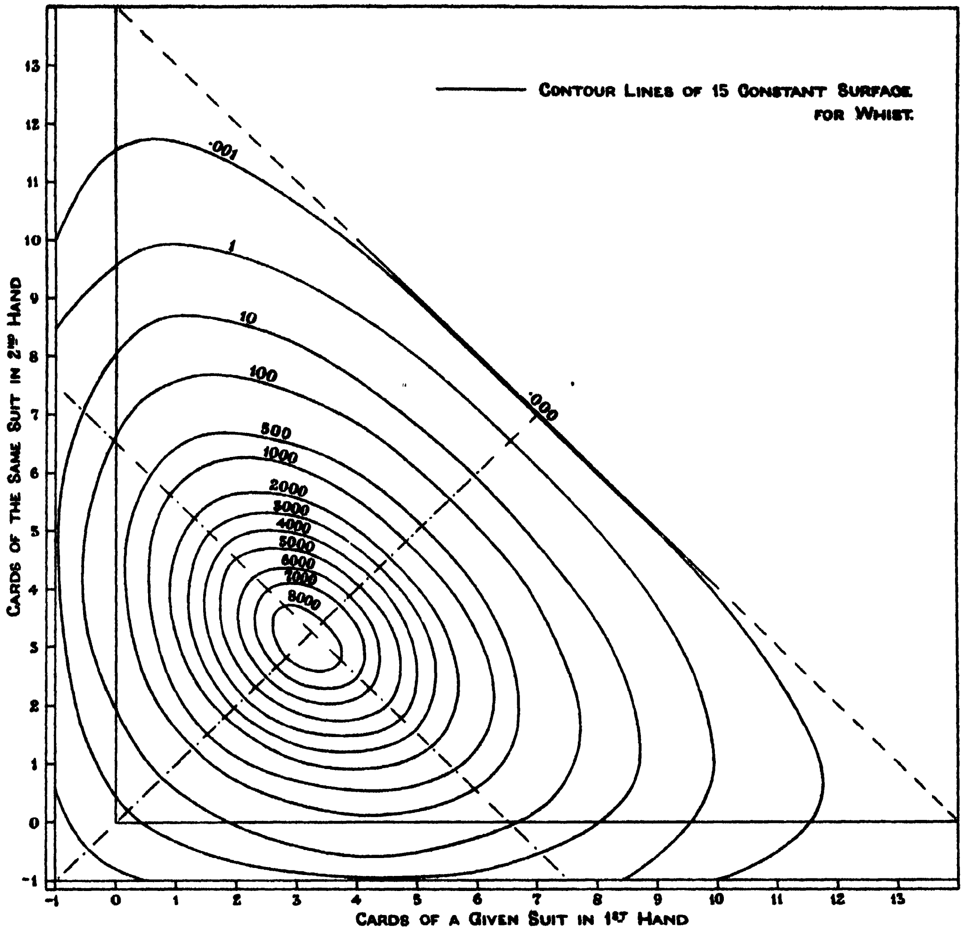


Diagram II.

The ordinates for this curve are given in Table III, p. 290, and compared with those of other solutions (*Biometrika*, Vol. XVI. pp. 180—1 *).

The superiority of the Type I curve over other representations is manifest. This manifestation in my experience always appears when an adequate number of categories are dealt with and a proper test of goodness of fit applied. It leads one to believe that if the corresponding frequency *surface* could be determined it would undoubtedly in like manner show its superiority over our present 15-constant surface.

It is further obvious that as far as the marginal totals are concerned the

---

* The frequencies are here calculated from the ordinates, as I have since realised that having treated the moments as discrete—i.e. without corrections—this was the more proper course to pursue. On p. 185 of the above memoir they are dealt with as areas and for totals of 1000, not 25,000. In Equation (xxiv), p. 182, for z_x the value 29150·101 is incorrectly given. This is the modal value of the ordinate, but must be replaced by $10^{39} \times$ ·824,7586 in the reduced form of the curve. The curve was worked from its logarithmic form and this error did not affect the computations. It was introduced when the non-logarithmic form of the curve was put down for publication, and the modal value of z_x inserted by analogy—of course a false one—with Equation (xxix), p. 183.

TABLE III.

Number of Cards of given suit	Theoretical Frequencies	Type I Curve	Tetrachoric Series	Marginal Totals Filon-Isserlis Surface	Marginal Totals Non-Skew Surface
0	319·8	311·4	341·0	$\left\{\begin{matrix} -1 & 19{\cdot}8 \\ 0 & 325{\cdot}1 \end{matrix}\right\}$ 344·9	$\left\{\begin{matrix} -2 & 4{\cdot}0 \\ -1 & 46{\cdot}5 \\ 0 & 426{\cdot}3 \end{matrix}\right\}$ 476·8
1	2001·6	1984·2	1956·4	1916·1	1931·7
2	5146·8	5192·2	5134·1	5140·4	4833·0
3	7158·2	7161·6	7218·1	7269·6	7089·8
4	5965·2	5928·4	5952·7	5979·6	6243·6
5	3117·3	3117·2	3095·4	3060·0	3275·4
6	1039·1	1052·4	1046·6	1020·7	982·8
7	220·4	221·9	223·7	229·1	155·3
8	29·2	27·3	28·6	35·3	11·1
9	2·3	1·7	2·1	3·3	0·3
10	0·1	·043	0·1	0·3	0·0
11	·0002	·0003	·00	·00	·00
12	·0000	·0000	·00	·00	·00
13	·0000	·0000	·00	·00	·00
χ^2, 0 to 10	—	1·262	3·270	12·11	155·65
P	—	·999	·972	·350	·000,000

15-constant surface gives far better results than either the Filon-Isserlis or the symmetrical surface. The conclusions to be drawn are essentially these:

(i) No curve of frequency is adequate which does not give β_1 and β_2 the same value for curve and data.

(ii) Of two frequency curves which both have the same β_1 and β_2 as the data, that which approximates to a convergent series will be better than that which approximates to a divergent series.

(9) *Illustration II.* I take as a second illustration the correlation table of the contemporaneous barometric heights at Southampton and Laudale. The data are given in Table IV (p. 291). The constants are as follows:

Southampton (x)	Laudale (y)
Mean $\bar{x} = 29''{\cdot}9839$	Mean $\bar{y} = 29''{\cdot}8488$.

The remaining constants are in tenths of inches as units:

$\sigma_x = 3{\cdot}250{,}067$, $\sigma_y = 3{\cdot}932{,}290$,

$p_{11} = 9{\cdot}971{,}435$, $p_{21} = 11{\cdot}919{,}404$, $p_{12} = 15{\cdot}598{,}613$,

$p_{31} = 361{\cdot}329{,}845$, $p_{22} = 399{\cdot}245{,}042$, $p_{13} = 500{\cdot}535{,}855$;

whence: $r = {\cdot}780{,}819$, $q_{21} = {\cdot}286{,}962$, $q_{12} = {\cdot}310{,}386$,

$q_{31} = 2{\cdot}676{,}586$, $q_{22} = 2{\cdot}444{,}352$, $q_{13} = 2{\cdot}532{,}831$.

TABLE IV.

Contemporaneous Barometric Heights at Southampton and Laudale.

→ + *x*. Southampton. Height in Inches (Central Values).

+ *y* ← Laudale. Height in Inches (Central Values).

	31·0	*30·9*	*30·8*	*30·7*	*30·6*	*30·5*	*30·4*	*30·3*	*30·2*	*30·1*	*30·0*	*29·9*	*29·8*	*29·7*	*29·6*	*29·5*	*29·4*	*29·3*	*29·2*	*29·1*	*29·0*	*28·9*	*28·8*	*28·7*	*28·6*	*28·5*	Totals
30·8	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
30·7	—	1	—	4·5	4	3	1·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	14
30·6	—	1	3	7·75	9·75	8·5	3·5	2·25	·25	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	36
30·5	—	1	·5	6·25	9·25	17	15·75	6·5	5·5	1·75	0·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	64
30·4	—	—	·5	5·75	11·75	22	24·75	39·75	19·75	12·25	4·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	141
30·3	—	—	—	5·25	11·5	15·5	24·75	55	50	25·5	6	5	1·25	·25	—	—	—	—	—	—	—	—	—	—	—	—	200
30·2	—	—	—	1	4·25	21·75	24	38·5	55·5	66	32	12·5	4·75	2·75	—	—	—	—	—	—	—	—	—	—	—	—	263
30·1	—	—	—	—	2	10·75	21·75	23·75	55·75	55·5	46·75	26	15	3·25	—	—	—	—	—	—	—	—	—	—	—	—	260·5
30·0	—	—	—	—	—	5	11	26·25	48·25	63	58	27·5	24·5	10	1·5	1·5	1	—	—	—	—	—	—	—	—	—	277·5
29·9	—	—	—	—	—	4	4	19	31·75	54·25	63·25	57	27·5	11·5	8·75	1	0·5	1	—	—	—	—	—	—	—	—	283·5
29·8	—	—	—	—	—	—	7	13	24·25	41·75	59·75	48	36·75	18	17·5	6·5	3·5	0·5	—	1	—	—	—	—	—	—	277·5
29·7	—	—	—	—	—	—	2	8	12	36·5	41·25	47	42·25	19·5	16	7·5	5	5	2·5	0·5	—	—	—	—	—	—	245
29·6	—	—	—	—	—	—	0·5	4	8·5	16	26·5	47	32·75	25·5	23·75	12·5	8	3	1	3	—	—	—	—	—	—	212
29·5	—	—	—	—	—	—	—	1	3	13	18	31	45	39·5	20	10·5	8	2·5	—	0·5	—	—	—	—	—	—	192
29·4	—	—	—	—	—	—	—	—	0·5	5	13·5	17·25	24·5	23·75	18	13·5	7·25	5·25	3	0·5	1	1	1	—	—	—	135
29·3	—	—	—	—	—	—	—	—	—	2	6	13·25	17·25	18·5	11·5	10·5	7·75	5·75	1	2	1	1	—	—	—	—	97·5
29·2	—	—	—	—	—	—	—	—	—	2	5·25	5·25	6	14	13	8	7	3	2	—	1	—	—	—	1	—	67·5
29·1	—	—	—	—	—	—	—	—	—	1	0·25	2·75	5·5	10·5	10	12·5	6	9·5	3	—	1	1	—	—	—	—	63
29·0	—	—	—	—	—	—	—	—	—	—	1	—	4	3	5·75	3·75	6·5	6	4·5	2·5	1·5	—	—	—	—	—	38·5
28·9	—	—	—	—	—	—	—	—	—	—	—	—	—	1	2·75	5·75	2·5	4·5	4	—	2	1	1	—	—	—	24·5
28·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	2·5	2	1·5	0·5	0·5	—	3	—	—	—	—	11
28·7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0·5	1·5	—	1·5	1	3	—	—	—	—	—	—	7·5
28·6	—	—	—	—	—	—	—	—	—	—	—	—	1	—	0·5	1	—	1	—	1	—	—	—	—	—	—	4·5
28·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
28·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0·5	—	—	—	1	—	—	1	2·5
28·3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0·5	—	—	—	—	—	—	—	·5
28·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
28·1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
28·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	1
27·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1
Totals	1	4	4	30·5	52·5	107·5	140·5	237	315	395·5	382·5	339·5	288	201	150·5	98·5	65	50	23·5	15·5	7·5	7	3	—	1	2	2922

For the marginal totals we have:

$$
\begin{aligned}
{}_x\beta_1 &= {\cdot}171{,}140, & {}_y\beta_1 &= {\cdot}224{,}535^+,\\
{}_x\beta_2 &= 3{\cdot}612{,}028, & {}_y\beta_2 &= 3{\cdot}194{,}947,\\
\sqrt{{}_x\beta_1} &= {\cdot}413{,}691, & \sqrt{{}_y\beta_1} &= {\cdot}473{,}852.
\end{aligned}
$$

The above lead to:

$$
\begin{aligned}
&B_2 = {\cdot}025{,}5012, \quad B_2' = {\cdot}008{,}1228,\\
&Q_{31} = {\cdot}055{,}6882, \quad Q_{22} = {\cdot}018{,}7496, \quad Q_{13} = {\cdot}031{,}7290.
\end{aligned}
$$

The total frequency $N = 2922$.

Hence by some further laborious arithmetic the fifteen constants of Equations (x)—(xv) and (xxxv)—(xliii) were determined and their values being substituted in Equation (i) there resulted the following surface:

$$
\begin{aligned}
z = 58{\cdot}243{,}994\, & e^{-1{\cdot}2809{,}9466\left(\frac{x^2}{\sigma_x^2} - 1{\cdot}561{,}638\frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)}\\
&\times\Big\{1{\cdot}137{,}4261 - {\cdot}198{,}0935\frac{x}{\sigma_x} - {\cdot}199{,}0118\frac{y}{\sigma_y}\\
&\quad - {\cdot}502{,}9167\frac{x^2}{\sigma_x^2} + {\cdot}531{,}4255\frac{xy}{\sigma_x\sigma_y} - {\cdot}186{,}8886\frac{y^2}{\sigma_y^2}\\
&+ {\cdot}234{,}3790\frac{x^3}{\sigma_x^3} - {\cdot}249{,}8810\frac{x^2y}{\sigma_x^2\sigma_y} - {\cdot}114{,}8185\frac{xy^2}{\sigma_x\sigma_y^2} + {\cdot}209{,}3993\frac{y^3}{\sigma_y^3}\\
&\quad \frac{x^4}{\sigma_x^4} - {\cdot}393.6955\frac{x^3y}{\sigma_x^3\sigma_y} + {\cdot}332{,}3922\frac{x^2y^2}{\sigma_x^2\sigma_y^2} - {\cdot}129{,}4956\frac{xy^3}{\sigma_x\sigma_y^3}\\
&\qquad + {\cdot}026{,}3067\frac{y^4}{\sigma_y^4}\Big\} \quad \ldots\ldots\text{(lxxii)}.
\end{aligned}
$$

Some remarks may be made on this equation before we proceed to the computing of the ordinates. We shall only calculate these to two decimal places; the numerical coefficients given are not correct to their last figure, we should have had to compute the original constants to more than six figures to obtain such accuracy. But they are sufficiently accurate to give us correct values of the ordinates to two decimal places, when we multiply up by large values of the higher powers. We have taken our axes of x and y in the usual directions, i.e. from left to right and downwards of the correlation table. Thus the values of x and y if positive denote the tenths of inches to be ***subtracted*** from the means of x and y to obtain the true barometric heights. The reader who desires x and y to give additive corrections to the means has only to change the sign of both x and y in the above equation for z. We may remark that there is no sign of convergency in the terms of the polynomial of successive orders. Such goodness of fit as the surface may present must depend, not on its being a convergent expansion, but on the conception that a surface of equal volume, with the same centroid and the fifteen momental coefficients the same as the data, cannot differ very widely from the form of the true surface.

It will be noticed that the mode of this surface as well as that of the previous illustration depends upon solving two simultaneous equations of the fifth order. The only way, as far as I can see, to obtain an accurate value of the position of the modal ordinate would be to solve these equations by successive approximation; but a fairly reasonable value can be found from the ordinates, when these have been calculated. This we shall now proceed to do.

Table V (p. 294) gives the central ordinates of the occupied cells of the original table. Although it is not correct to consider the ordinate of the cell as a measure of its frequency, it gives a rough appreciation and the comparison with Table IV compels us to adopt the view that the original data are very ragged, not impossibly heterogeneous*, and that it would be difficult for any continuous mathematical surface to represent them; all it can do is to graduate them with more or less success†.

In Table V are given the calculated ordinates of the 15-constant surface. But the labour of calculating 450 ordinates was very great, and the deduction of the corresponding volumes (Table VI, p. 295) added to the work. It was found to be too toilsome to determine the amount of frequency in the extreme "marches" of the 15-constant surface. The total of the volumes showed that the uncomputed tails of the arrays must contain some 13 units in all (actually 13·03) of frequency. These 13 units were distributed in 93 cells of the "marches" in the manner indicated in brackets in Table VI. It will be seen that in no case was an additional unit placed in any cell, the average of these added cells being ·14 of a unit. The calculation of an additional 100 ordinates was thus avoided. The distribution was of course a rough appreciation, and was done before the marginal totals had been calculated by the tetrachoric expansions. The amounts added to the arrays are in each case in the decimal places, and until tables of the bivariate tetrachoric functions are available some such appreciative process of distributing the frequency in the marches will have to be adopted. No statistician can be called upon to calculate 500 or 600 ordinates and the corresponding volumes.

From the ordinates of the 15-constant surface (in Table V) the contour lines were found by a system of sections and the work was, if laborious, relatively easy; the system of curves in Diagram III was obtained. We next proceeded to deal with the data (Table IV) themselves. The volumes were replaced by the ordinates

---

* We ought probably, in the case of two barometric stations, to deal separately with the correlations for the winter and summer months, possibly we ought to use only three month periods. There are signs of at least two modes in the present correlation table.

† Table VII, p. 298, shows the ordinates as found from the observed frequencies (volumes) by aid of Formula (π) of the Appendix. Although this treatment of the frequencies has itself a slight smoothing effect the appearance of negative ordinates creates some trouble when the series of sections are plotted. It is from this table that ultimately the contours of the observation-surface were determined. The reader will appreciate not only the difficulty of proceeding from such a series of irregular ordinates to the contours, but how much room is left for arbitrary steps in the smoothing. Too much stress cannot therefore be laid on the deviation between the observational and the 15-constant surface contours. It would have been by no means difficult to admit a bias to a greater agreement to play its part, but this was carefully guarded against. The reader may rest assured that if the contours are an individual solution of the problem, they are a solution which is absolutely independent of the 15-constant surface contours.

TABLE V.

Computed Ordinates of 15-Constant Surface fitted to Barometric Data.

Southampton. Heights in Inches.

Laudale. Heights in Inches.

y	*31·1*	*31·0*	*30·9*	*30·8*	*30·7*	*30·6*	*30·5*	*30·4*	*30·3*	*30·2*	*30·1*	*30·0*	*29·9*	*29·8*	*29·7*	*29·6*	*29·5*	*29·4*	*29·3*	*29·2*	*29·1*	*29·0*	*28·9*	*28·8*	*28·7*	*28·6*
	(·02)	(·02)	(·03)	(·03)	(·01)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
31·0	·07	·10	·14	·15	·05	−·14	−·28	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·9	·12	·20	·36	·62	·74	·51	·04	−·29	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·8	·14	·27	·57	1·19	1·94	2·17	1·50	·44	−·08	·00	(·00)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·7	·14	·29	·66	1·61	3·29	4·85	4·89	3·18	1·09	−·07	−·28	(·00)	(·00)	—	—	—	—	—	—	—	—	—	—	—	—	—
30·6	·12	·27	·61	1·66	4·12	7·68	10·09	9·10	5·42	1·89	·17	−·16	−·07	(·00)	(·01)	—	—	—	—	—	—	—	—	—	—	—
30·5	·09	·21	·49	1·36	3·98	9·25	15·37	17·87	14·38	7·80	2·68	·50	·04	·03	·03	(·01)	—	—	—	—	—	—	—	—	—	—
30·4	·05	·15	·35	·94	3·05	8·70	18·12	26·56	27·32	19·62	9·73	3·31	·86	·27	·13	·06	(·02)	—	—	—	—	—	—	—	—	—
30·3	—	·09	·24	·58	1·89	6·45	16·82	30·92	39·88	36·28	23·32	10·69	3·72	1·20	·50	·19	·07	(·01)	—	—	—	—	—	—	—	—
30·2	—	—	·15	·34	1·00	3·82	12·39	28·57	46·02	52·21	42·07	24·36	10·56	3·82	1·54	·65	·24	·07	(·01)	—	—	—	—	—	—	—
30·1	—	—	·08	·20	·50	1·86	7·29	21·12	42·51	59·84	59·67	42·72	22·60	9·43	3·95	1·73	·71	·23	·06	(·00)	—	—	—	—	—	—
30·0	—	—	—	·11	·27	·82	3·50	12·88	31·69	55·34	67·98	59·62	38·18	18·57	8·38	3·84	1·74	·66	·19	·04	(·00)	(·00)	—	—	—	—
29·9	—	—	—	·06	·15	·38	1·46	6·16	19·27	41·72	63·02	67·44	52·11	29·71	14·80	7·15	3·54	1·56	·54	·14	·02	·01	—	—	—	—
29·8	—	—	—	—	·08	·20	·62	2·62	9·78	26·02	48·16	62·73	58·38	39·19	21·87	11·21	6·02	3·05	1·25	·39	·09	·02	(·00)	—	—	—
29·7	—	—	—	—	—	·10	·29	1·09	4·36	13·78	30·90	48·74	54·53	43·21	27·31	14·91	8·52	4·92	2·41	·90	·25	·05	·01	—	—	—
29·6	—	—	—	—	—	·05	·14	·48	1·86	6·49	17·13	32·32	43·27	40·54	29·31	17·13	10·18	6·56	3·81	1·72	·58	·14	·03	—	—	—
29·5	—	—	—	—	—	—	·06	·22	·81	2·89	8·55	18·84	29·88	33·09	27·64	17·50	10·53	7·25	4·96	2·72	1·11	·33	·07	·01	(·00)	—
29·4	—	—	—	—	—	—	·02	·09	·35	1·26	3·99	9·96	18·42	24·07	23·45	16·52	9·99	6·87	5·12	3·55	1·77	·64	·17	·03	·01	(·00)
29·3	—	—	—	—	—	—	—	·03	·14	·52	1·72	4·86	10·32	15·85	18·16	14·73	9·31	6·01	4·87	3·85	2·34	1·03	·34	·08	·01	(·00)
29·2	—	—	—	—	—	—	—	·01	·05	·20	·57	2·18	5·27	9·45	12·80	12·32	8·75	5·42	4·07	3·54	2·58	1·39	·54	·15	·03	(·01)
29·1	—	—	—	—	—	—	—	—	·01	·06	·17	·88	2·41	5·04	8·06	9·34	7·88	5·22	3·50	2·94	2·44	1·57	·74	·25	·06	(·02)
29·0	—	—	—	—	—	—	—	—	—	·02	·04	·31	·96	2·35	4·44	6·19	6 35	4·92	3·28	2·44	2·06	1·52	·85	·35	·11	(·03)
28·9	—	—	—	—	—	—	—	—	—	(·00)	·01	·09	·33	·94	2·09	3·50	4·37	4·10	3·05	2·14	1·67	1·31	·84	·41	·15	(·04)
28·8	—	—	—	—	—	—	—	—	—	—	(·00)	·02	·10	·32	·83	1·66	2·51	2·87	2·53	1·88	1·37	1·05	·73	·41	·17	(·04)
28·7	—	—	—	—	—	—	—	—	—	—	—	—	·03	·09	·28	·66	1·20	1·65	1·75	1·48	1·11	·82	·59	·36	·18	(·05)
28·6	—	—	—	—	—	—	—	—	—	—	—	—	·01	·02	·09	·22	·47	·78	1·00	1·00	·82	·62	·44	·29	·16	(·04)
28·5	—	—	—	—	—	—	—	—	—	—	—	—	(·00)	(·00)	·02	·06	·16	·31	·47	·56	·53	·43	·31	·21	·13	(·03)
28·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·00)	·01	·04	·10	·18	·25	·28	·26	·20	·14	·09	(·02)
28·3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·01)	·03	·06	·10	·13	·13	·12	·09	·06	(·02)
28·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·01)	·02	·03	·05	·06	·06	·05	·03	(·01)
28·1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·00)	·01	·02	·02	·02	·02	·02	(·00)
28·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·00	·00	·01	·01	·01	(·00)	—
27·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·00)	(·00)	·00	·00	(·00)	—	—

Values in brackets have been used in computing certain volumes.

TABLE VI.

Frequencies computed from Ordinates of 15-*Constant Surface fitted to Barometric Data.*

Height of Barometer at Southampton in Inches.

	31·1	31·0	30·9	30·8	30·7	30·6	30·5	30·4	30·3	30·2	30·1	30·0	29·9	29·8	29·7	29·6	29·5	29·4	29·3	29·2	29·1	29·0	28·9	28·8	28·7	28·6	Totals
31·0	(·00)	(·08)	(·10)	(·14)	(·16)	(·10)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·58
30·9	(·01)	(·10)	·36	·62	·75	·54	(·20)	(·04)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2·62
30·8	(·02)	·20	·58	1·19	1·93	2·18	1·56	(·60)	(·05)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	8·31
30·7	(·04)	·27	·68	1·63	3·26	4·79	4·89	3·29	(·40)	(·20)	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	19·45
30·6	(·05)	·29	·64	1·70	4·13	7·58	9·95	9·10	5·62	2·13	(·20)	(·20)	(·10)	—	—	—	—	—	—	—	—	—	—	—	—	—	41·69
30·5	(·15)	(·27)	·51	1·43	4·06	9·20	15·12	17·62	14·41	8·10	2·99	·66	(·40)	(·10)	—	—	—	—	—	—	—	—	—	—	—	—	75·02
30·4	(·09)	(·21)	·37	1·01	3·19	8·78	17·91	26·07	26·95	19·72	10·14	3·66	1·02	(·30)	(·10)	—	—	—	—	—	—	—	—	—	—	—	119·52
30·3	(·04)	(·15)	(·24)	·63	2·04	6·67	16·84	30·43	39·10	35·86	23·54	11·18	4·07	1·35	·55	(·40)	(·10)	—	—	—	—	—	—	—	—	—	173·19
30·2	—	—	(·10)	·36	1·11	4·08	12·67	28·41	45·16	51·19	41·70	24·71	11·07	4·13	1·66	·70	(·50)	(·10)	—	—	—	—	—	—	—	—	227·65
30·1	—	—	—	(·25)	·56	2·07	7·69	21·39	42·04	58·62	58·59	42·52	23·01	9·89	4·17	1·82	·76	(·50)	(·11)	—	—	—	—	—	—	—	273·99
30·0	—	—	—	(·15)	·29	·93	3·85	13·33	31·82	54·51	66·57	58·70	38·18	19·04	8·70	3·99	1·81	·71	(·30)	(·10)	—	—	—	—	—	—	302·98
29·9	—	—	—	—	(·20)	·43	1·66	6·64	19·77	41·58	61·91	66·11	51·50	29·95	15·13	7·35	3·64	1·62	·58	(·30)	(·05)	—	—	—	—	—	308·42
29·8	—	—	—	—	(·05)	(·30)	·71	2·92	10·32	26·40	47·74	61·57	57·35	39·04	22·08	11·42	6·11	3·11	1·31	·42	·10	(·05)	—	—	—	—	291·00
29·7	—	—	—	—	—	(·15)	·32	1·23	4·74	14·30	31·07	48·14	53·52	42·74	27·31	15·10	8·60	4·96	2·46	·95	·28	(·10)	—	—	—	—	255·97
29·6	—	—	—	—	—	(·10)	(·25)	·54	2·06	6·89	17·53	32·26	42·62	39·99	29·12	17·27	10·26	6·56	3·83	1·77	·62	·16	(·05)	—	—	—	211·88
29·5	—	—	—	—	—	(·04)	(·10)	·24	·90	3·12	8·91	19·05	29·63	32·67	27·34	17·57	10·65	7·25	4·92	2·74	1·15	·36	(·15)	—	—	—	166·79
29·4	—	—	—	—	—	—	(·05)	(·15)	·39	1·37	4·22	10·22	18·44	23·84	23·15	16·50	10·12	6·91	5·11	3·52	1·75	·67	·19	·04	—	—	126·64
29·3	—	—	—	—	—	—	—	(·04)	·16	·57	1·85	5·06	10·45	15·79	17·92	14·62	9·40	6·11	4·85	3·80	2·33	1·05	·36	·09	(·05)	—	94·50
29·2	—	—	—	—	—	—	—	—	(·10)	·22	·65	2·30	5·40	9·50	12·67	12·17	8·75	5·52	4·11	3·51	2·56	1·40	·56	·16	(·10)	—	69·68
29·1	—	—	—	—	—	—	—	—	—	(·10)	·21	·94	2·51	5·13	8·03	9·22	7·80	5·25	3·56	2·95	2·42	1·56	·75	·26	(·20)	—	50·89
29·0	—	—	—	—	—	—	—	—	—	—	(·10)	·34	1·03	2·43	4·48	6·14	6·27	4·89	3·31	2·47	2·04	1·51	·85	·36	(·20)	—	36·42
28·9	—	—	—	—	—	—	—	—	—	—	—	(·20)	·36	1·00	2·15	3·51	4·33	4·05	3·04	2·16	1·68	1·30	·84	·41	(·25)	(·05)	25·33
28·8	—	—	—	—	—	—	—	—	—	—	—	(·05)	(·20)	·35	·87	1·70	2·51	2·84	2·51	1·88	1·38	1·05	·73	·41	(·25)	(·05)	16·78
28·7	—	—	—	—	—	—	—	—	—	—	—	—	(·05)	·10	·30	·69	1·22	1·65	1·74	1·47	1·11	·82	·59	·36	(·20)	(·05)	10·35
28·6	—	—	—	—	—	—	—	—	—	—	—	—	—	·02	·10	·24	·49	·79	1·00	·99	·82	·62	·44	·29	(·15)	(·05)	6·00
28·5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·04)	·07	·17	·32	·48	·56	·53	·43	·31	·21	(·13)	(·03)	3·28
28·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·04)	(·10)	(·16)	·19	·25	·28	·26	·20	·14	(·12)	(·02)	1·76
28·3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·02)	(·08)	(·09)	·10	·13	·13	·12	·09	(·05)	—	·81
28·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·03)	(·05)	(·05)	·05	·06	·06	·05	—	—	·35
28·1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	(·01)	(·01)	(·02)	·02	·02	·02	·02	—	—	·12
28·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·00	·01	·01	·01	—	—	·03
Totals	·40	1·57	3·58	9·11	21·73	47·85	93·86	162·04	243·99	324·88	377·92	387·87	350·91	277·36	205·87	140·52	93·61	63·42	43·56	30·01	19·30	11·56	6·23	2·90	1·70	·25	2922

Volumes in brackets are the estimated distribution of 13 units in the "marches."

calculated from the formulae in the Appendix. These are given in Table VII, p. 298. Then these ordinates were plotted as sections, and the real trouble began*. Although the computation of the ordinates from nine adjacent volumes in itself exerted a smoothing effect, yet the section-curves remained erratic, being often bimodal. Whether this was to be attributed to random sampling or to heterogeneity in the material, it is not possible at present to say. When attempting many years ago a general theory of skew correlation, I found the contour lines of a number of observational frequency surfaces, but found them by no means easy to reduce to a mathematically useful smoothness. It has always been a matter of admiration to me that Galton should have deduced the system of concentric, similarly

CONTOURS OF 15 CONSTANT SURFACE
FOR BAROMETRIC HEIGHTS AT SOUTHAMPTON & LAUDALE

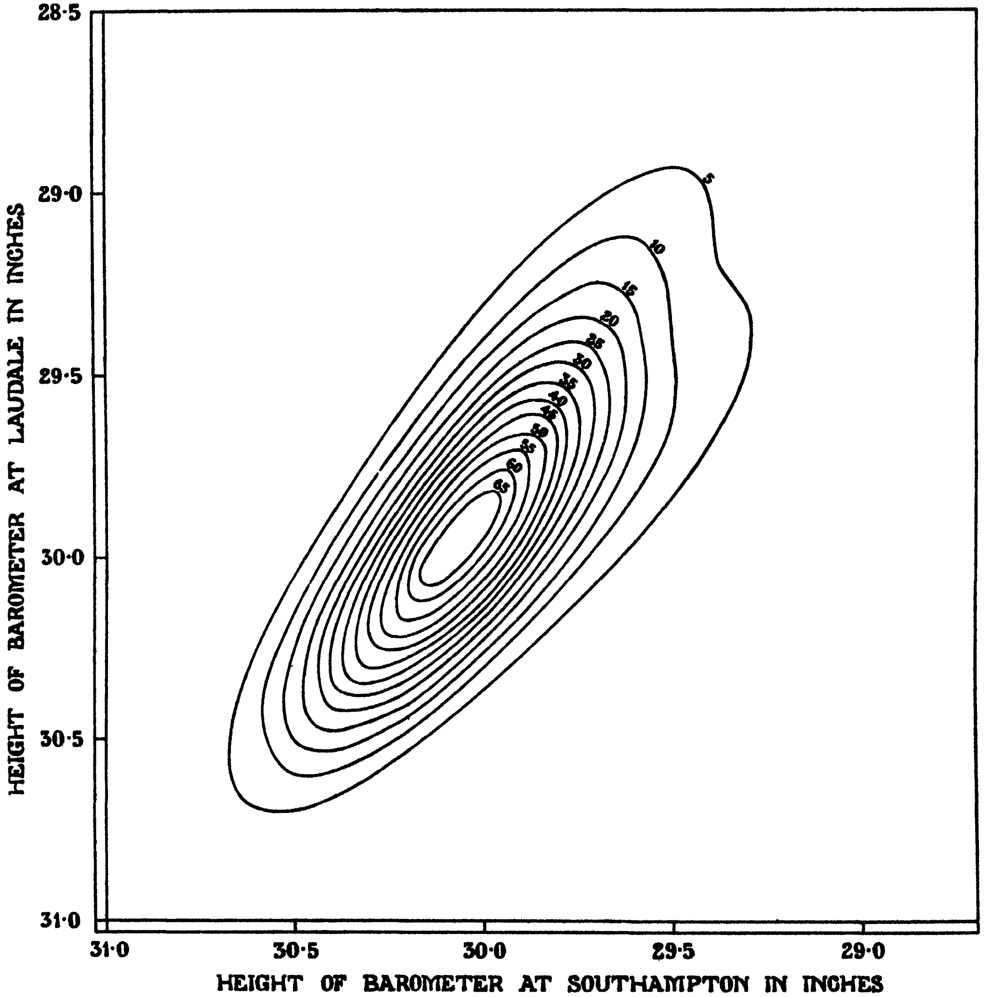


Diagram III.

---

* See footnote, p. 298.

placed and similar ellipses of the normal surface from a table of some 350 pairs corresponding to stature in Father and Son. Here we have nearly 3000 pairs of variates, but fail to reach anything like so great regularity as he attained on his far inferior numbers. In the case of the sections the curves were drawn by aid of the spline, without smoothing out double modes and other irregularities. The contours were plotted from these sections, and these contours were then smoothed, there being an individual drawing of each. They were occasionally asymmetrical lemniscatoid curves! The irregularities were gradually smoothed out, and the contours placed one on top of the other, being properly orientated. The contours were now once more adjusted to each other. All this was done without paying any attention to the 15-constant surface's contours. The result is seen in Diagram IV.

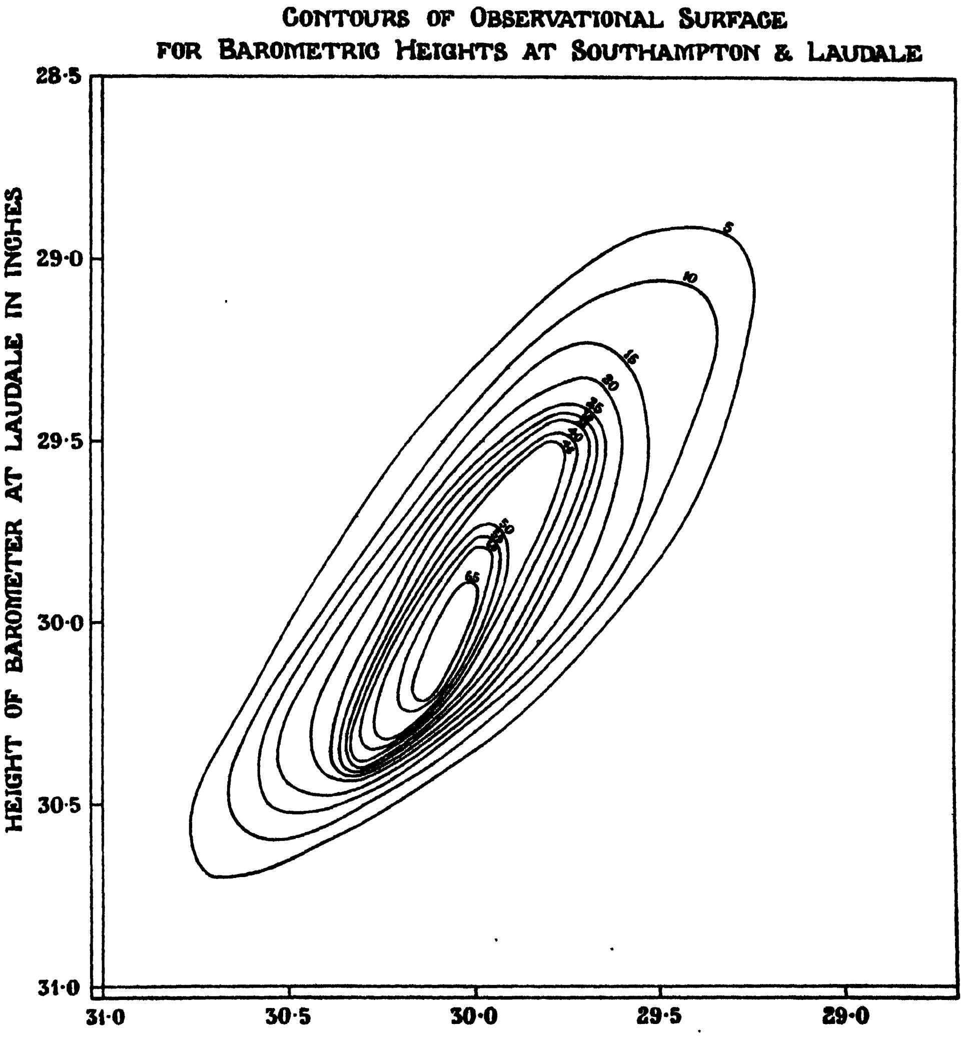


Diagram IV.

TABLE VII.

Ordinates from Observed Frequencies of Barometric Heights.

Southampton. Heights of Barometer in Inches.

	31·0	30·9	30·8	30·7	30·6	30·5	30·4	30·3	30·2	30·1	30·0	29·9	29·8	29·7	29·6	29·5	29·4	29·3	29·2	29·1	29·0	28·9	28·8	28·7	28·6
31·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·9	−·04	−·04	·00	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·8	1·13	1·08	−·03	−·20	−·17	−·14	−·96	·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·7	−·04	1·09	−·37	4·77	3·94	2·91	1·49	−·15	−·01	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·6	−·04	·95	3·13	8·06	10·16	8·53	2·89	2·22	−·04	−·08	−·02	—	—	—	—	—	—	—	—	—	—	—	—	—	—
30·5	−·04	1·11	·13	6·32	8·91	17·53	16·27	4·86	5·27	1·27	·33	−·02	—	—	—	—	—	—	—	—	—	—	—	—	—
30·4	−·05	−·67	·33	5·70	11·68	22·81	24·59	42·03	18·48	12·16	4·49	−·44	−·05$^-$	−·01	—	—	—	—	—	—	—	—	—	—	—
30·3	—	·0	−·25	5·37	11·93	14·69	23·88	57·86	51·92	24·06	4·17	5·05$^-$	1·04	·12	−·01	—	—	—	—	—	—	—	—	—	—
30·2	—	·0	−·04	·77	3·41	23·17	23·53	38·31	55·92	70·13	31·83	11·73	4·21	2·88	−·12	—	—	—	—	—	—	—	—	—	—
30·1	—	·0	·00	−·13	1·72	10·40	22·55	21·69	57·50	55·02	47·41	26·13	15·06	2·62	−·19	−·11	−·04	—	—	—	—	—	—	—	—
30·0	—	—	—	·00	−·29	4·75	10·44	26·40	48·91	64·53	59·35	25·07	25·31	9·99	·87	1·62	1·09	−·09	—	—	—	—	—	—	—
29·9	—	—	—	—	−·18	4·32	2·92	19·07	30·93	54·98	64·21	59·70	26·60	10·73	8·91	·44	·31	1·14	−·04	−·04	—	—	—	—	—
29·8	—	—	—	—	·01	−·51	7·41	12·71	24·22	41·37	61·69	47·59	37·24	17·44	18·38	6·35$^-$	3·57	·17	−·17	1·16	−·04	—	—	—	—
29·7	—	—	—	—	—	−·08	1·68	8·06	10·73	38·03	41·01	47·37	43·69	18·48	15·81	7·09	4·82	5·39	2·65$^+$	·30	−·02	—	—	—	—
29·6	—	—	—	—	—	−·09	·33	3·91	8·50	15·09	25·80	49·22	31·46	24·93	24·67	12·51	8·15$^-$	2·80	·80	3·43	−·13	—	—	—	—
29·5	—	—	—	—	—	·00	−·06	·87	2·53	13·45$^-$	17·48	30·81	47·24	41·37	19·46	9·98	8·17	2·23	−·30	·44	−·06	−·04	—	—	—
29·4	—	—	—	—	—	·00	−·09	−·06	·25$^+$	4·61	13·86	16·68	24·27	23·50	18·16	13·84	7·00	5·36	3·23	·30	1·07	1·04	−·04	−·04	—
29·3	—	—	—	—	—	—	—	—	−·10	1·79	5·56	13·57	17·54	18·82	10·89	10·56	7·77	6·01	·61	2·25$^-$	·95$^+$	1·09	1·13	·00	—
29·2	—	—	—	—	—	—	—	—	−·08	1·99	5·59	4·98	5·22	14·34	13·38	7·51	7·16	2·46	2·06	−·22	1·09	−·09	−·09	−·04	−·04
29·1	—	—	—	—	—	—	—	—	−·04	1·08	−·10	2·76	5·44	10·91	9·91	13·47	5·48	10·37	2·82	−·28	1·02	1·13	·00	·00	1·17
29·0	—	—	—	—	—	—	—	—	·0	−·09	1·17	−·36	4·34	2·59	5·92	3·06	6·86	5·94	4·61	2·68	1·52	−·15$^+$	−·04	·00	−·04
28·9	—	—	—	—	—	—	—	—	—	·00	−·04	·01	−·22	·93	2·63	6·26	2·11	4·67	4·28	−·39	2·24	·91	−·04	−·04	·00
28·8	—	—	—	—	—	—	—	—	—	—	—	·00	·00	−·08	·93	2·48	2·08	1·39	·29	·44	−·24	3·48	1·13	·00	·00
28·7	—	—	—	—	—	—	—	—	—	—	—	—	−·04	−·02	·46	1·59	−·22	1·61	·95$^+$	3·41	−·13	−·14	−·18	·00	·00
28·6	—	—	—	—	—	—	—	—	—	—	—	—	1·17	−·07	·52	1·08	−·09	1·11	−·13	1·04	−·04	·00	·01	·00	·00
28·5	—	—	—	—	—	—	—	—	—	—	—	—	−·04	·00	−·02	−·04	—	−·04	·02	−·04	·00	·00	−·04	·00	·00
28·4	—	—	—	—	—	—	—	—	—	—	—	—	·00	—	·00	·00	—	−·02	·56	−·02	·00	−·04	1·17	−·04	−·04
28·3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−·02	·56	−·02	·00	·00	−·04	—	·00
28·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·00	−·02	·00	·00	·00	·00	—	·00
28·1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	·00	−·04	·00	—	—	—	·00
28·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−·04	1·17	−·04	—	—	—	·00
27·9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	−·04	·00	—	—	—	−·04

The observational surface is seen to be very precipitate as we pass from the centroid* in a direction S.E. by S., and north-west of the centroid is a table-land. Here possibly there ought to be a second low barometer mode, if there be really heterogeneity in the material. It may be asked why, with this possibility before us, did we select barometric material? The answer is that—as first noticed by Quételet—the barometric frequency at a given station forms a practically important case of definite skew variation, that the marginal frequencies of a barometric table—or the frequencies of heights at single stations—do not exhibit definite traces of bimodalism; and last but not least that this particular correlation table was one for which all the 15 constants had been determined for other purposes, and so the surface could be obtained without excessive labour. Owing to the "precipice" it was not possible to superpose Diagrams III and IV in one drawing. Tissues of III and IV are provided in the pocket at the end of this volume, and the reader will easily convince himself of the goodness of fit of the contour lines in the case of I and II and the badness of fit in the case of III and IV. But then he must remember that in the case of I and II, both are mathematical surfaces, while in the case of III and IV, III is a mathematical surface and IV is a smoothed series of observations, themselves subject to all the irregularities of random sampling. It does not necessarily follow that the 15-constant surface is a bad graduation of the data. A study of the data in the table suggests that no continuous mathematical surface is likely to give a first-class result.

In view of the good success of the 15-constant surface in the case of whist, I do not feel wholly discouraged by the bad fit of the contours for the barometric data. We have not at present enough comparative material to be able to say what is really good or bad in fitting theoretical frequency contours to observational contours.

Even before passing judgment in this case it is well to study other aspects of the fitted surface. I will deal first with the regression lines.

If x be the Southampton barometric height, y that of Laudale, $\tilde{y}_x$ the mean of the heights at Laudale for a given height x at Southampton, and $\tilde{x}_y$ the mean of the heights at Southampton for a given height y at Laudale, the origin of all quantities being the mean heights 29″·9839 at Southampton and 29″·8488 at Laudale, we have by Equation (lv) for the regression curves:

$$\tilde{y}_x = 3{\cdot}932{,}290 \quad {\cdot}780{,}819\,\frac{x}{\sigma_1} - \frac{{\cdot}018{,}0280\left(\dfrac{x^2}{\sigma_1^2} - 1\right) + {\cdot}023{,}9590\left(\dfrac{x^3}{\sigma_1^3} - 3\,\dfrac{x}{\sigma_1}\right)}{1 + {\cdot}068{,}9485\left(\dfrac{x^3}{\sigma_1^3} - 3\,\dfrac{x}{\sigma_1}\right) + {\cdot}025{,}5012\left(\dfrac{x^4}{\sigma_1^4} - 6\,\dfrac{x^2}{\sigma_1^2} + 3\right)} \quad \text{......(lxxiii)},$$

$$3{\cdot}250{,}067\left\{{\cdot}780{,}819\,\frac{y}{\sigma_2} - \frac{{\cdot}029{,}8035\left(\dfrac{y^2}{\sigma_2^2} - 1\right) - {\cdot}006{,}3593\left(\dfrac{y^3}{\sigma_2^3} - 3\,\dfrac{y}{\sigma_2}\right)}{1 + {\cdot}078{,}9753\left(\dfrac{y^3}{\sigma_2^3} - 3\,\dfrac{y}{\sigma_2}\right) + {\cdot}008{,}1228\left(\dfrac{y^4}{\sigma_2^4} - 6\,\dfrac{y^2}{\sigma_2^2} + 3\right)}\right. \quad \text{......(lxxiv)}.$$

---

* By an oversight the centroid has been omitted in Diagrams III and IV, but the reader will easily insert it from the values of its coordinates on p. 290, i.e. Southampton 29″·984, Laudale 29″·849.

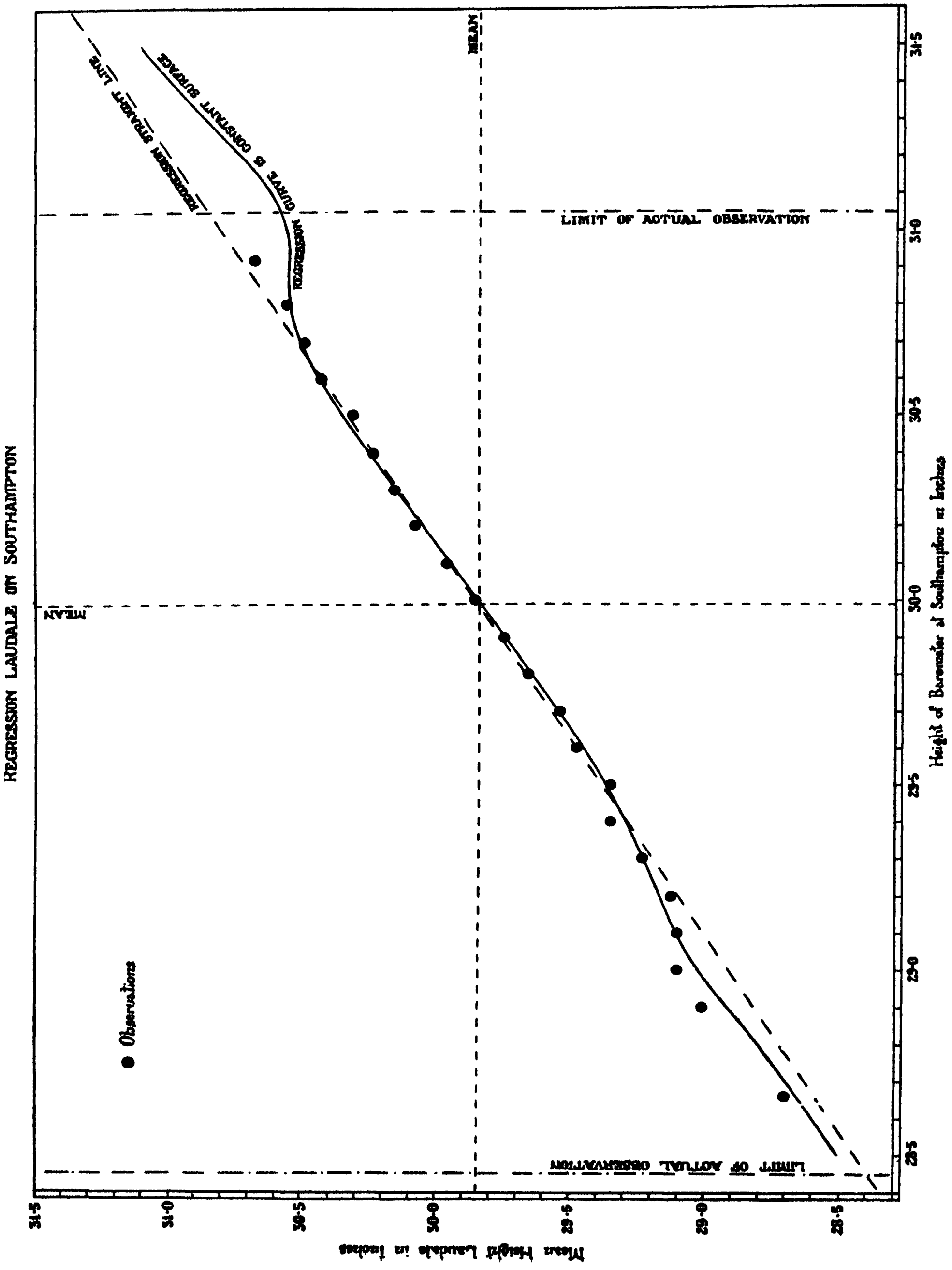

REGRESSION LAUDALE ON SOUTHAMPTON
MEAN
MEAN
LIMIT OF ACTUAL OBSERVATION
LIMIT OF ACTUAL OBSERVATION
REGRESSION STRAIGHT LINE
REGRESSION CURVE IS CONSTANT SURFACE
● Observations
28·5
29·0
29·5
30·0
30·5
31·0
31·5
Height of Barometer at Southampton in Inches
28·5
29·0
29·5
30·0
30·5
31·0
31·5
Mean Height Laudale in Inches

Diagram V.

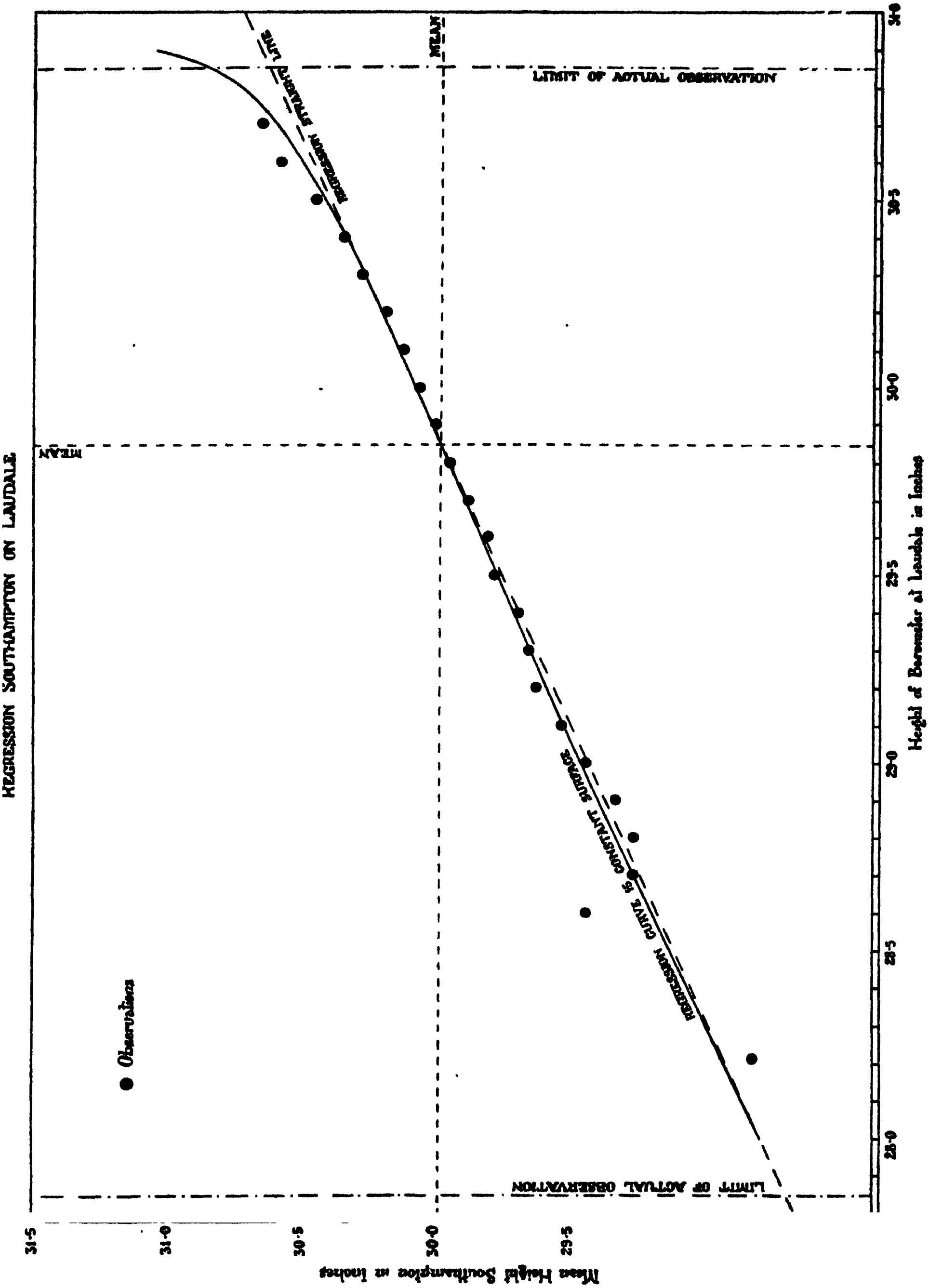

REGRESSION SOUTHAMPTON ON LAUDALE
MEAN
MEAN
LIMIT OF ACTUAL OBSERVATION
LIMIT OF ACTUAL OBSERVATION
REGRESSION STRAIGHT LINE
REGRESSION CURVE to CONSTANT SURFACE
● Observations
Mean Height Southampton in Inches
31·5
31·0
30·5
30·0
29·5
Height of Barometer at Laudale in Inches
28·0
28·5
29·0
29·5
30·0
30·5
31·0

Diagram VI.

The labour of computing $\tilde{y}_x$ and $\tilde{x}_y$ was considerable, but was accomplished for each $\frac{1}{10}''$ of barometric height. The results are plotted in Diagrams V and VI, which show the observations, the regression curves of the 15-constant surface and the regression straight lines. However algebraically complicated the regression curves, and geometrically quaint their graphs, there is small doubt that they fit very well, and possibly better than the regression straight lines, the observation points.

I showed in 1897* that Type III frequency curves fitted to barometric data gave a physical limit to high barometer at each station. This limit was $31''\cdot511$ at Laudale. The marginal total frequency curve at Laudale becomes *negative* at $30''\cdot972$. This is shown by the numerator in the second part of the value of $\tilde{x}_y$ becoming zero. It follows that the regression quintic is asymptotic to the line $y = 30''\cdot972$. It is shown turning up toward this line in Diagram VI. The asymptote naturally lies at a height outside experience, the maximum observed height at Laudale being $30''\cdot796$. No similar asymptote was found in the case of the Southampton marginal frequency†.

The reader is asked to consider carefully the significance of the fact that these regression curves have not been fitted by determining their five constants so as to give the best fit of the curves to the array means. The array means are not directly used in their determination. They are, so to speak, a bye-product of the 15-constant surface and depend on the observational moments and product moments. It is this indirectness of fitting which makes their goodness of fit—within, of course, the limits of observation—really a strong point for the 15-constant surface. Outside the limits of observation they may in this case and probably in others act eccentrically. They cannot be said to have a strong theoretical foundation, but granted the equality of the 15 constants, then they are giving a quite reasonable result. Knowing how closely the regression was linear (the two η's and r being nearly equal) it was a surprise to find how well the quintics bent to their tasks!

I will now turn to the marginal totals. We have from Equations (xlvi) and (xlvii):

$$\int_{-\infty}^{x} z_x dx = N\{{}_{x'}\tau_0 - \cdot068{,}9485\,{}_{x'}\tau_3' - \cdot025{,}5012\,{}_{x'}\tau_4'\} \quad \ldots\ldots\ldots\text{(lxxv)},$$

$$\int_{-\infty}^{y} z_y dy = N\{{}_{y'}\tau_0 - \cdot078{,}9753\,{}_{y'}\tau_3' - \cdot008{,}1228\,{}_{y'}\tau_4'\} \quad \ldots\ldots\text{(lxxvi)},$$

and the fundamental property $\int \tau_s dx = -\frac{1}{\sqrt{s}}\,\tau_{s-1}(x)$.

Here x' and y' stand for x/σ_1 and y/σ_2, ${}_{x'}\tau_0$ and ${}_{y'}\tau_0$ for the corresponding probability integrals, i.e. $\frac{1}{2}(1 + \alpha_x)$ and $\frac{1}{2}(1 + \alpha_y)$ of the usual notation, while

$${}_{x'}\tau_3' = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}(x'^2 - 1), \qquad {}_{x'}\tau_4' = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x'^2}(x'^3 - 3x'),$$

$${}_{y'}\tau_3' = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y'^2}(y'^2 - 1), \qquad {}_{y'}\tau_4' = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y'^2}(y'^3 - 3y').$$

---

* *Phil. Trans.* Vol. 190, A, pp. 435, 446—8.

† The reader will notice that on the low barometer side the quintics are asymptotic to the respective regression straight lines.

Unfortunately the *Tables for Statisticians and Biometricians*, Part I, give the tetrachoric functions for argument $\frac{1}{2}(1-\alpha_x)$ and not for x, and so are not best suited to our present purpose*. We have been obliged therefore to use other tables. Perhaps the most convenient are those of Professor James W. Glover, entitled: *Tables of Applied Mathematics in Finance, Insurance, Statistics*, Ann Arbor, 1923.

He tables the derivatives of $\phi(t)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$ as $\phi^2(t)$, $\phi^3(t)$, $\phi^4(t)$, etc. The relation between the integral of τ_s and Glover's ϕ^s is:

$$\int \tau_s d\left(\frac{x}{\sigma_1}\right)=(-1)^{s-1}\frac{1}{\sqrt{s!}}\phi^{s-2}(t),$$

where $t=x/\sigma_1$. Accordingly we find for Southampton:

$$\int_{-\infty}^{t} z_x dx = 2922\left\{\tfrac{1}{2}(1+\alpha_t)-{}\cdot068{,}9485\,\phi^2(t)+{}\cdot025{,}5012\,\phi^3(t)\right\},$$

and for Laudale, if $t=y/\sigma_2$:

$$\int_{-\infty}^{t} z_y dy = 2922\left\{\tfrac{1}{2}(1+\alpha_t)-{}\cdot078{,}9753\,\phi^2(t)+{}\cdot008{,}1228\,\phi^3(t)\right\},$$

where we must remember that:

(*a*) for t *negative*, e.g. barometer for Southampton above mean 29″·9839, we must use $\frac{1}{2}(1-\alpha_t)$ for $\frac{1}{2}(1+\alpha_t)$ and further:

(*b*) for t *negative* $\phi^3(t)$ must have the *opposite* sign to its tabled value. $\phi^2(t)$ for t negative retains the same sign as in Glover's *Tables*.

If these points be borne in mind it is straightforward, but laborious, to compute the probability integrals of the marginal total, and by subtraction of their values to obtain the cell frequencies. The following table gives the results with comparison of the margins for Rhodes' surface.

Mr E. S. Pearson for Mr Rhodes' groups finds the values $P = {}\cdot59$ and $\cdot62$ for the Southampton and Laudale frequencies fitted with Pearson skew curves. The results do not give one the confidence we might desire in the tetrachoric series for a skew univariate frequency. The tetrachoric gives a good result for Southampton, where the Rhodes' surface marginal total is poor; it gives a poor result for Laudale, where the fit of the Rhodes' surface margin is passable. One would have thought that the equalising of the four moment coefficients would have given a better result than the equalising of only three. It certainly does so, when Pearson skew curves are used. I have reworked independently Equation (lxxvi) and recomputed the cell frequencies, but only to reach the same result. The chief region of error is from height 29″·9 to height 30″·4, but I can see no reason for the fall here. We are bound I think to conclude that although it is better to equalise four than three moments, the form of the curve fitted to the frequency may count for a good

---

* The Second Part of the *Tables for Statisticians* will contain a full table for the tetrachoric functions to x as argument in our form. See also Dr Alice Lee's Table in the present issue, pp. 351—4.

TABLE VIII.

Marginal Barometric Table Totals.

Barometric Height	Southampton Observed	Southampton Computed Tetrachoric	Southampton Computed Rhodes	Laudale Observed	Laudale Computed Tetrachoric	Laudale Computed Rhodes
31·0	1	3·1	0·1	—		
30·9	4 } 9	3·6 } 15·5	1·2 } 6·9	—	7·2	6·1
30·8	4	8·8	5·6	2		
30·7	30·5	21·2	19·3	14	19·9	15·2
30·6	52·5	47·2	50·7	36	41·4	36·8
30·5	107·5	92·9	105·1	64	74·7	71·6
30·4	140·5	161·2	178·9	141	119·6	124·0
30·3	237	244·0	257·9	200	173·0	181·3
30·2	315	324·1	322·0	263	227·7	236·4
30·1	395·5	377·5	359·9	260·5	274·0	278·6
30·0	382·5	387·6	359·4	277·5	302·9	300·6
29·9	339·5	350·4	327·5	283·5	308·9	302·0
29·8	288	282·4	275·3	277·5	291·9	284·0
29·7	201	206·2	215·1	245	256·8	251·6
29·6	150·5	140·4	157·7	212	212·6	212·8
29·5	98·5	93·4	108·9	192	167·4	171·6
29·4	65	63·0	71·4	135	127·4	132·0
29·3	50	43·6	44·7	97·5	94·8	98·8
29·2	23·5	29·8	26·4	67·5	69·8	71·3
29·1	15·5	19·3	15·0	63	50·9	50·0
29·0	7·5	11·5	8·5	38·5	36·4	34·0
28·9	7	6·2	4·5	24·5	25·2	22·6
28·8	3			11	16·6	14·7
28·7	— } 6	4·9	3·8	7·5	10·4	9·2
28·6	1			4·5		
28·5	2			—		
28·4	—	—	—	2·5		
28·3	—	—	—	0·5	12·4	13·4
28·2	—	—	—	—		
28·1	—	—	—	—		
28·0	—	—	—	1		
27·9	—	—	—	1		
Totals	2922	2922	2918·8	2922	2922	2918·8
	$\chi^2(n'=21)$	= 18·024	= 29·942	$\chi^2(n'=23)$	= 39·086	= 24·466
	P	= ·586	= ·071†	P	= ·087	= ·325*

deal. It is not possible to discuss directly the marginal curve of the Rhodes' surface; it is not yet expressed in a finite form, but it appears to be an expansion in terms of Pearson's Type III Curve and its differential coefficients, as the tetrachoric expansion is of the normal curve and its differential coefficients†. Its fitness for describing univariate frequencies might well be discussed.

---

* Mr Rhodes gives the values ·2 and ·4 for P, using 19 groups for Southampton and 21 for Laudale (*Biometrika*, Vol. XIV. p. 374). Using his groups I find for Southampton: $P = \cdot 110$ and for Laudale: $P = \cdot 295$.

† See Romanovsky, *Biometrika*, Vol. XVI. p. 114 *et seq.*

The following table indicates the values of the marginal totals as found (i) from the tetrachoric expansion and (ii) from the volumes as deduced by the formulae of the Appendix from the ordinates (Table V) together with the distribution made for the marches (see Table VI).

TABLE IX.

Barometric Height in Inches.	SOUTHAMPTON				LAUDALE			
	Computed Main Surface	Appreciated Marches	Totals	Tetrachoric Expansion	Computed Main Surfaces	Appreciated Marches	Totals	Tetrachoric Expansion
31·1	—	·40	·40	3·1	—	—	—	—
31·0	·76	·81	1·57		-	·58	·58	−0·5
30·9	3·14	·44	3·58	3·6	2·27	·35	2·62	
30·8	8·57	·54	9·11	8·8	7·64	·67	8·31	7·8
30·7	21·32	·41	21·73	21·2	18·81	·64	19·45	19·9
30·6	47·25	·60	47·85	47·2	41·14	·55	41·69	41·4
30·5	93·17	·69	93·86	92 9	74·10	·92	75·02	74·7
30·4	161·21	·83	162·04	161·2	118·82	·70	119·52	119·6
30·3	243·44	·55	243·99	244·0	172·26	·93	173·19	173·0
30·2	342·58	·30	324·68	324·1	226·95	·70	227·65	227·6
30·1	377·62	·30	377·92	377·5	273·13	·86	273·99	274·0
30·0	387·42	·45	387·87	387·6	302·43	·55	302·98	302·9
29·9	350·16	·75	350·91	350·4	307·87	·55	308·42	308·9
29·8	276·96	·40	277·36	282·4	290·60	·40	291·00	291·9
29·7	205·73	·14	205·87	206·2	255·72	·25	255·97	256·8
29·6	140·08	·44	140·52	140·4	211·48	·40	211·88	212·6
29·5	92·89	·72	93·61	93·4	166·50	·29	166·79	167·4
29·4	62·54	·88	63·42	63·0	126·44	·20	126·64	127·4
29·3	43·00	·56	43·56	43·6	94·41	·09	94·50	94·8
29·2	29·54	·47	30·01	29·8	69·48	·20	69·68	69·8
29·1	19·25	·05	19·30	19·3	50·59	·30	50·89	50·9
29·0	11·41	·15	11·56	11·5	36·12	·30	36·42	36·4
28·9	6·03	·20	6·23	6·2	24·83	·50	25·33	25·2
28·8	2·90	·00	2·90	4·9	16·23	·55	16·78	16·6
28·7	—	1·70	1·70		10·05	·30	10·35	10·4
28·6	—	·25	·25		5·80	·20	6·00	6·1
28·5	—	—	—	—	3·08	·20	3·28	6·4
28·4	—	—	-	—	1·32	·44	1·76	
28·3	—	—	—	—	·57	·24	·81	
28·2	—		-	—	·22	·13	·35	
28·1	—	—	—	—	·08	·04	·12	
28·0	—	-	—	—	·03	·00	·03	
Totals	2908·97	13·03	2922	2922	2908·97	13·03	2922	2922

I think we must conclude from these results that our method of computing volumes from ordinates is in the main very satisfactory. The small disagreement at the tops of the columns is largely due to our not computing and then allowing for the negative ordinates of the 15-constant surface, while they appear in the tetrachoric marginal totals.

General Goodness of Fit of the 15-*Constant Surface.* We may now consider the problem of fit apart from the contour lines and the marginal totals, i.e. by the P, χ^2 test for goodness of fit. Unluckily our goodness of fit tables (*Tables for*

Statisticians and Biometricians, Table XII, p. 28) are limited at present to 30 cells*, whereas for a thorough test in this matter it would be desirable to work with 80 to 100 frequency groups at least, but the numerical work thus involved cannot at present be faced. Accordingly I was reluctantly compelled to reduce the cell frequencies to 30 or less. If this must be done, I concluded that it might be just as well to use the 24 groups adopted by Mr Rhodes, as to do so would enable me to compare the general efficiency of his skew surface with the present one. Mr Rhodes' surface has nine available constants. One of these goes in determining the volume; two others in making his origin the mode. He is thus left with six constants which may be determined from σ_1, σ_2, ${}_x\beta_1$, ${}_y\beta_1$, ${}_x\beta_2$, ${}_y\beta_2$. It is clear therefore that q_{21}, q_{12} and, I think, r are functions of these marginal constants, or the correlational constants even to the second and third orders are not independent of the marginal variations†. The 15-constant surface which leaves absolute freedom to these constants ought therefore to give a better fit than Mr Rhodes' surface.

---

* This was fully adequate in the case of univariate frequency for which the tables were originally computed.

† Using our notation for the β's we have with Mr Rhodes' symbols θ, ϕ, λ:

$$\frac{2\,{}_x\beta_2 - 3\,{}_x\beta_1 - 6}{8\,{}_x\beta_1} = \frac{\lambda\{\phi(\phi-1)\}^2}{(\phi^3+\lambda)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(a),$$

$$\frac{2\,{}_y\beta_2 - 3\,{}_y\beta_1 - 6}{3\,{}_y\beta_1} = \frac{\lambda\{\theta(\theta-1)\}^2}{(\theta^3+\lambda)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(b),$$

$$\frac{{}_x\beta_1}{{}_y\beta_1} = \frac{(\phi^3+\lambda)^2\,(\theta^2+\lambda)^3}{(\phi^2+\lambda)^3\,(\theta^3+\lambda)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(c).$$

These three equations theoretically suffice to determine θ, ϕ and λ in terms of the marginal variation constants, ${}_x\beta_1$, ${}_x\beta_2$, ${}_y\beta_1$ and ${}_y\beta_2$.

But

$$r = \frac{(\theta\phi+\lambda)}{\sqrt{(\theta^2+\lambda)(\phi^2+\lambda)}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(d),$$

$$q_{21} = r\sqrt{{}_x\beta_1} + \sqrt{\tfrac{1}{3}(1-r^2)(2\,{}_x\beta_2 - 3\,{}_x\beta_1 - 6)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(e),$$

$$q_{12} = r\sqrt{{}_y\beta_1} + \sqrt{\tfrac{1}{3}(1-r^2)(2\,{}_y\beta_2 - 3\,{}_y\beta_1 - 6)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(f).$$

Thus not only r but q_{21} and q_{12} are functions of the β's of the marginal totals, and the same objections hold as in the case of the Filon-Isserlis surface. Actually Mr Rhodes pays no attention to fourth moments and would pass over Equations (a) and (b) above, deducing his values of θ, ϕ and λ from (c) and (e) and (f) by replacing $2\,{}_x\beta_2 - 3\,{}_x\beta_1 - 6$ in those equations by the values for it in (a) and (b); thus indirectly the ${}_x\beta_2$ and ${}_y\beta_2$ are involved. I do not know what values should be given to the sign of the second radicals in (e) and (f), but if we substitute the values of r and the β's observed for Southampton and Laudale, we find

$$q_{21} = \cdot 323{,}018 \pm \cdot 304{,}071,$$
$$= \cdot 018{,}947 \text{ or } \cdot 627{,}089,$$

either of which is very far from the observed value ·286,962. On the other hand $2\,{}_y\beta_2 - 3\,{}_y\beta_1 - 6$ is negative or the value of q_{12} is imaginary. Hence neither of the conditions connecting q_{21} and q_{12} with the β's of the marginal totals (and these might be treated as criteria justifying the use of the surface) is even approximately satisfied by the Southampton-Laudale data. If we assume that q_{21} and q_{12} have their observed values, then we find from the Rhodes' surface:

$${}_x\beta_2 = 3{\cdot}261{,}705 \text{ and } {}_y\beta_2 = 3{\cdot}350{,}457$$

against the observed values:

Southampton 3·612,028 and Laudale 3·194,947.

The divergence is quite sufficient to account for the inferiority in the fit of the marginal totals when we compare the P-values with those of curves having both β_2's equalised with the β_2's of the data (see *Biometrika*, Vol. XIV. p. 374). At the same time the probable errors of these β_2's are only about half the differences between the observed and computed values. One great advantage of Mr Rhodes' surface lies in the relative ease with which the ordinates and the corresponding frequencies can be calculated.

The accompanying table provides the frequencies of the 15-constant surface, of the Rhodes' surface and of the data in 24 groups.

TABLE X.

Abridged Table showing the Correlated Data for Barometric Heights at Southampton and Laudale in 24 Groups.

Height of Barometer at Southampton in Inches.

<table>
<tr><th></th><th></th><th>—30·7</th><th>30·6—30·4</th><th>30·3—30·1</th><th>30·0—29·8</th><th>29·7—29·5</th><th>29·4—29·2</th><th>29·1—</th></tr>
<tr><td rowspan="8">Height of Barometer at Laudale in Inches.</td><td>—30·6</td><td>O. 19·25
T. 19·0
R. 16·9</td><td>O. 30·25
T. 44·8
R. 36·6</td><td rowspan="2">O. 218·5
T. 189·3
R. 170·4</td><td rowspan="2">O. 17·25
T. 22·9
R. 19·1</td><td rowspan="3" colspan="3">O. 20·25
T. 26·9
R. 24·1</td></tr>
<tr><td>30·5—30·3</td><td colspan="2">O. 171·5
T. 162·9
R. 191·7</td></tr>
<tr><td>30·2—30·0</td><td rowspan="2" colspan="2">O. 118·5
T. 111·8
R. 115·5</td><td>O. 432·5
T. 450·2
R. 462·8</td><td>O. 247·0
T. 231·3
R. 223·9</td></tr>
<tr><td>29·9—29·7</td><td>O. 240·5
T. 257·8
R. 262·0</td><td>O. 422·75
T. 449·9
R. 422·5</td><td>O. 106·25
T. 116·7
R. 130·1</td><td rowspan="2" colspan="2">O. 64·5
T. 64·0
R. 58·3</td></tr>
<tr><td>29·6—29·4</td><td colspan="3">O. 51·5
T. 46·9
R. 42·7</td><td>O. 255·5
T. 248·7
R. 239·5</td><td>O. 187·0
T. 162·0
R. 188·7</td></tr>
<tr><td>29·3—29·1</td><td rowspan="3" colspan="4">O. 72·5
T. 67·2
R. 59·2</td><td>O. 108·5
T. 100·6
R. 107·0</td><td>O. 45·0
T. 39·7
R. 49·5</td><td>O. 8·0
T. 14·2
R. 9·2</td></tr>
<tr><td>29·0—28·8</td><td rowspan="2">O. 29·0
T. 35·3
R. 32·4</td><td rowspan="2">O. 36·5
T. 37·2
R. 38·3</td><td>O. 11·5
T. 13·4
R. 10·0</td></tr>
<tr><td>28·7—</td><td>O. 8·0
T. 9·0
R. 8·4</td></tr>
</table>

O. = Observed Frequency; T. = 15-Constant Surface Frequency;
R. = Rhodes' Surface Frequency.

Totals: O. = 2922; T. = 2921·7; R. = 2918·8;
χ^2 (24 cells): for T. = 29·092; for R. = 36·328;
P: for T. = ·1775; for R. = ·0646*.

It will be seen that the 15-constant surface has three to four times (if we accept Mr Rhodes' $P = \cdot 04$) the goodness of fit of Rhodes' surface. I am inclined to think the fit would have been substantially better had the blocking up of the central cells not been so extensive. In the cell Southampton 29''·7—29''·5, Laudale 29''·6—29''·4, there is possibly some slip in either the 15-constant surface or the Rhodes' surface value. I have checked the former without discovering any error.

---

* Mr Rhodes gives (*Biometrika*, Vol. XIV. p. 375) the value ·04 for P. My value somewhat increases his, but my arithmetic has not been checked.

The agreement of the 15-constant surface with the data is not as good as one might have desired, but it is probably all that could have been anticipated having regard to the nature of the data. If the 15-constant surface actually describes the material sampled, we should expect a worse result than that observed in one out of five or six trials. Such a result is sufficiently encouraging to make it worth while investigating the 15-constant surface on further material. It is, for reasons already stated, not an ideal solution of the bivariate frequency surface, but it is better than anything thus far attempted. Its freedom, as far as all momental constants up to the fourth order are concerned, renders it much more efficient than any surface can be which has arbitrary relations, between the fourth and third order momental constants.

(10) *On a practically adequate Method of finding the Mode of the* 15-*Constant Surface.* I will illustrate this on the present surface for the barometric data.

After the ordinates have been tabled we pick out the three arrays one way in which the mode must lie; in our case we have:

Southampton.

Laudale.	*30''·2*	*30''·1*	*30''·0*	Scheme of Ordinates		
30''·2	52·21	—	—	$z_{2,2}$	—	—
30''·1	59·84	59·67	—	$z_{2,1}$	$z_{1,1}$	—
30''·0	55·34	67·98	59·62	$z_{2,0}$	$z_{1,0}$	$z_{0,0}$
29''·9	—	63·02	67·44		$z_{1,9}$	$z_{0,9}$
29''·8	—	—	62·73	—	—	$z_{0,8}$

We now determine the position of the vertex of the parabola passing through each triplet of ordinates. If we have three ordinates equally spaced, z_{-1}, z_0, z_{+1}, then the distance $\tilde{y}$ of abscissa corresponding to vertex from the foot of central ordinate* is, for three ordinates z_{+1}, z_0, z_{-1}:

$$\tilde{y} = \frac{1}{2}\frac{z_{+1} - z_{-1}}{2z_0 - z_{+1} - z_{-1}},$$

and the maximum ordinate

$$\tilde{z} = z_0 + \frac{1}{8}\frac{(z_{+1} - z_{-1})^2}{2z_0 - z_{+1} - z_{-1}} = \frac{\left\{z_0 - \left(\frac{\sqrt{z_{+1}} + \sqrt{z_{-1}}}{2}\right)^2\right\}\left\{z_0 - \left(\frac{\sqrt{z_{+1}} - \sqrt{z_{-1}}}{2}\right)^2\right\}}{z_0 - \frac{z_{+1} + z_{-1}}{2}}.$$

We now apply these formulae to the above three triplets and determine the mode in each case. Remembering that our unit is $\frac{1}{10}''$ we find for position of Laudale modes:

---

* In the old days I found these formulae quite useful in determining the position of the meridian and the latitude from three circum-meridian observations of the altitudes of the sun with the sextant.

Southampton.

	30″·2	*30″·1*	*30″·0*
Laudale Mode:	30″·08710	29″·98745	29″·88759
Modal Frequency:	59·94095	68·08571	67·53649

Now it will be seen that the Laudale modes for the three constant values at Southampton are nearly on a straight line; for practical purposes it will suffice to take them on a line at distances ·141245 apart. Accordingly we proceed to find the mode of the above three modes in the same manner; and we find it to be at a point on the line joining the three above modes between the second and third at distance from the second ·43683 of the interval between them. This gives for the position of the maximum ordinate[*]:

Southampton: 30″·0563, Laudale: 29″·9439,

while the modal ordinate is 68·91519.

Working with the Laudale arrays instead of those for Southampton I found:

Southampton: 30″·0579, Laudale: 29″·9483,

and the modal ordinate 68·89727.

The means from both determinations:

Southampton: 30″·0571, Laudale: 29″·9461,

and the modal ordinate 68·9062, will I believe be very close to the true values for the 15-constant surface.

(11) *Conclusions.* While I believe expansions in tetrachoric functions, whether univariate or bivariate, have neither from the standpoint of the theory of probability nor from the theory of functions any demonstrable validity—for in the case of frequency expansions they do not usually converge—they have some practical value depending upon the fact that any two expressions for which some of the momental constants are equal, will show more or less agreement between their values, according to the number of momental constants equalised. Yet if two systems be taken to represent a frequency distribution by equalising the same number of momental constants, one may be much better than the other—the one that is ultimately based on a converging series will give the better result when a proper goodness of fit test is applied. Many curves can be propounded which will within a given range give four or more momental constants equal to those of a given frequency distribution, but I have not found any yet which on the whole approach frequency distributions with the same good accuracy of fit as the skew curves I have proposed[†]. They certainly give far better results than tetrachoric expansions taken up to the same number of equalised moment coefficients. I feel

---

[*] Mr Rhodes gives Southampton 30″·10845, Laudale 29″·99315 and modal ordinate 64·87118.

[†] Some of the comparisons made between my skew curves and tetrachoric expansions are not worth discussing; they consist in splitting the frequency up into five or six groups and making no real test of the tails at all. A parabola of the fifth or sixth order could be made to fit such groupings even more successfully!

quite confident that if a surface could be determined of the same character as the curves with the same number of free momental constants, it would be far more satisfactory than an expansion in bivariate tetrachoric functions. But after many years of investigation this has not been achieved, and for the present we are thrown back on expansions by bivariate tetrachoric functions. The present paper shows that:

(*a*) The 15-constant tetrachoric expansion gives as judged by the contour lines an excellent expression for a mathematical bivariate distribution such as the whist double hypergeometrical series.

(*b*) For an actual series of observations it appears less successful, but this series is a case wherein there is some suspicion of heterogeneity, and after all its goodness of fit as measured by $P = \cdot 18$ means that one out of every five to six such series would give a worse result. This is about three times as good as the Rhodes' surface result.

It appears to me therefore that until something much better is forthcoming the tetrachoric bivariate expansion must be used.

It will certainly give much worse results than have been obtained in this paper, if we use it equating only the volume N and the constants $\bar{x}$, $\bar{y}$, σ_1, σ_2, r, ${}_1\beta_1$, ${}_2\beta_1$, q_{12} and q_{21}. As a univariate distribution requires for reasonable fit all the momental constants up to the fourth, so a bivariate distribution requires also all the momental constants up to the fourth, i.e. we require a 15-constant surface.

But the present work has demonstrated how great is the labour involved. It is not only the computing of all the moments up to the fourth of the data and correcting them for grouping; this is arduous enough. But when the numerical form of the surface has been obtained, it is practically of no value until: (*a*) the ordinates have been computed, and (*b*) from these ordinates the theoretical cell contents have been deduced.

After the experience we have had with this paper we can hardly recommend the process to statisticians, the labour is far too strenuous. But taking this paper to show that the contour-system of the 15-constant surface is remarkably flexible, and capable of giving something approximating to a fit, if not of the best, is there any way in which help can be given to the statistician?—Clearly there is. He could be largely aided by a computation of the bivariate tetrachoric functions exhibited in Equation (xlv). There are ten functions involved here and if we extract the factor $\dfrac{N}{\sigma_1\sigma_2}$, they depend solely on $x' = \dfrac{x}{\sigma_1}$, $y' = \dfrac{y}{\sigma_2}$ and r. These functions are

$$\tilde{z},\ \frac{d^3\tilde{z}}{dx'^3},\ \frac{d^3\tilde{z}}{dx'^2dy'},\ \frac{d^3z}{dx'dy'^2},\ \frac{d^3\tilde{z}}{dy'^3},$$

and
$$\frac{d^4\tilde{z}}{dx'^4},\ \frac{d^4\tilde{z}}{dx'^3dy'},\ \frac{d^4\tilde{z}}{dx'^2dy'^2},\ \frac{d^4\tilde{z}}{dx'dy'^3},\ \frac{d^4\tilde{z}}{dy'^4}.$$

Probably a range of x from 0 to $\pm 4\sigma_1$* by $\frac{1}{10}\sigma_1$ positive and negative, with a like range for y, and one for r from -1 to $+1$ by ·05 would be adequate, and clearly only six functions need be computed. This means a table of $6 \times (41)^3$, or 413,526, entries, or supposing a page of 7 columns of 60 entries each, a volume of 985 pages. The computation would not be hard if $\bar{z}$ were once calculated, especially if it could be cooperatively organised. But would there be any chance of finding the funds needful for publication?—I fear not, until there is greater recognition of the importance of statistical research. It is clear that the tabulation of $\bar{z}$ in itself would be of much value. At first sight it might appear as if volumes in the x and y argument ranges would be more valuable than the ordinates. But on consideration it will be seen that these are unlikely to coincide with the subranges of the actual data. I believe it would be best to table ordinates, to work from Equation (xlv) to the ordinates of the 15-constant surface at the required points, and then by some formulae similar to those of my Appendix, deduce the cell frequencies. While the computation might, as I have said, be possible by cooperation, it is the cost of publication which will block the way. Nevertheless, I cannot believe, that except in very special inquiries, the isolated statistician will feel able to give the labour and the time needful to fit a 15-constant surface to a skew-correlation table until tables of the bivariate tetrachoric functions are available. That is, I think, where we must leave the matter for the present.

While the general theory of the 15-constant surface was given by me in the lectures of last session, I have to express my obligation to Mr Wishart for checking my algebra; to Miss E. M. Elderton and Miss M. Moul I owe a deep debt for their services in calculating ordinates and cell frequencies, while Miss McLearn has struggled most valiantly in the production of the contours, those for the observed barometric data being of a most difficult, trying and uncertain character, which only her skill and patience have to some extent surmounted.

APPENDIX.

On Methods of proceeding from Cell Frequencies to Ordinates, or from Ordinates to Cell Frequencies, in the case of a bivariate Frequency Surface.

The Lagrangian interpolation formula which corresponds to the midpanel central difference formula up to and including second order differences takes the following form†:

$$\begin{aligned} z = z_{0,0}(1-x^2)(1-y^2) &- \tfrac{1}{2}y(1-y)(1-x^2)z_{0,-1} + \tfrac{1}{2}y(1+y)(1-x^2)z_{0,1} \\ &- \tfrac{1}{2}x(1-x)(1-y^2)z_{-1,0} + \tfrac{1}{2}x(1+x)(1-y^2)z_{1,0} \\ &+ \tfrac{1}{4}xy(1+x)(1+y)z_{1,1} + \tfrac{1}{4}xy(1-x)(1-y)z_{-1,-1} \\ &- \tfrac{1}{4}xy(1+x)(1-y)z_{1,-1} - \tfrac{1}{4}xy(1-x)(1+y)z_{-1,1}. \end{aligned}$$

---

* In a Table of three or four thousand entries we shall certainly need to go as far as $4\sigma_1$, possibly further.

† K. Pearson, *Tracts for Computers*, No. III, p. 26. Cambridge University Press.

This surface passes through the tops of all nine ordinates of the cell scheme:

		$-y$		
	$z_{-1,-1}$	$z_{0,-1}$	$z_{1,-1}$	
$-x$	$z_{-1,0}$	$z_{0,0}$	$z_{1,0}$	$+x$
	$z_{-1,1}$	$z_{0,1}$	$z_{1,1}$	
		$+y$		

Central cell ordinates.

		$-y$		
	$f_{-1,-1}$	$f_{0,-1}$	$f_{1,-1}$	
$-x$	$f_{-1,0}$	$f_{0,0}$	$f_{1,0}$	$+x$
	$f_{-1,1}$	$f_{0,1}$	$f_{1,1}$	
		$+y$		

Cell frequencies.

Corresponding to the cell ordinates z, we have the nine cell frequencies f. If we note the following integrals:

$$\int_{-\frac{1}{2}}^{+\frac{1}{2}}(1-x^2)\,dx=\int_{-\frac{1}{2}}^{+\frac{1}{2}}(1-y^2)\,dy=\tfrac{11}{12}; \qquad \int_{-\frac{1}{2}}^{+\frac{1}{2}}x(1-x)\,dx=\int_{-\frac{1}{2}}^{+\frac{1}{2}}y(1-y)\,dy=-\tfrac{1}{12};$$

$$\int_{-\frac{1}{2}}^{+\frac{1}{2}}x(1+x)\,dx=\int_{-\frac{1}{2}}^{+\frac{1}{2}}y(1+y)\,dy=\tfrac{1}{12}; \qquad \int_{+\frac{1}{2}}^{+\frac{3}{2}}(1-x^2)\,dy=\int_{+\frac{1}{2}}^{+\frac{3}{2}}(1-y^2)\,dy=-\tfrac{1}{12};$$

$$\int_{+\frac{1}{2}}^{+\frac{3}{2}}y(1-x)\,dx=\int_{+\frac{1}{2}}^{+\frac{3}{2}}y(1+y)\,dy=-\tfrac{1}{12}; \qquad \int_{+\frac{1}{2}}^{+\frac{3}{2}}x(1+x)\,dx=\int_{+\frac{1}{2}}^{+\frac{3}{2}}y(1+y)\,dy=\tfrac{25}{12};$$

we can readily integrate for the volume of every cell content, and find at once:

$$f_{0,0}=\tfrac{1}{576}\{484z_{0,0}+22(z_{0,1}+z_{0,-1}+z_{1,0}+z_{-1,0})+z_{1,1}+z_{1,-1}+z_{-1,1}+z_{-1,-1}\}\ \ldots(\alpha),$$

$$f_{1,1}=\tfrac{1}{576}\{4z_{0,0}-50z_{0,1}-2z_{0,-1}-50z_{1,0}-2z_{-1,0}+625z_{1,1}+25z_{1,-1}+25z_{-1,1}+z_{-1,-1}\}\ \ldots\ldots(\beta),$$

$$f_{1,-1}=\tfrac{1}{576}\{4z_{0,0}-2z_{0,1}-50z_{0,-1}-50z_{1,0}-2z_{-1,0}+25z_{1,1}+625z_{1,-1}+z_{-1,1}+25z_{-1,-1}\}\ \ldots\ldots(\gamma),$$

$$f_{-1,1}=\tfrac{1}{576}\{4z_{0,0}-50z_{0,1}-2z_{0,-1}-2z_{1,0}-50z_{-1,0}+25z_{1,1}+z_{1,-1}+625z_{-1,1}+25z_{-1,-1}\}\ \ldots\ldots(\delta),$$

$$f_{-1,-1}=\tfrac{1}{576}\{4z_{0,0}-2z_{0,1}-50z_{0,-1}-2z_{1,0}-50z_{-1,0}+z_{1,1}+25z_{1,-1}+25z_{-1,1}+625z_{-1,-1}\}\ \ldots\ldots(\epsilon),$$

$$f_{0,1}=\tfrac{1}{576}\{-44z_{0,0}+550z_{0,1}+22z_{0,-1}-2z_{1,0}-2z_{-1,0}+25z_{1,1}+z_{1,-1}+25z_{-1,1}+z_{-1,-1}\}\ \ldots\ldots(\zeta),$$

$$f_{0,-1}=\tfrac{1}{576}\{-44z_{0,0}+22z_{0,1}+550z_{0,-1}-2z_{1,0}-2z_{-1,0}+z_{1,1}+25z_{1,-1}+z_{-1,1}+25z_{-1,-1}\}\ \ldots\ldots(\eta),$$

$$f_{1,0}=\tfrac{1}{576}\{-44z_{0,0}-2z_{0,1}-2z_{0,-1}+550z_{1,0}+22z_{-1,0}+25z_{1,1}+25z_{1,-1}+z_{-1,1}+z_{-1,-1}\}\ \ldots\ldots(\theta),$$

$$f_{-1,0}=\tfrac{1}{576}\{-44z_{0,0}-2z_{0,1}-2z_{0,-1}+22z_{1,0}+550z_{-1,0}+z_{1,1}+z_{1,-1}+25z_{-1,1}+25z_{-1,-1}\}\ \ldots\ldots(\iota).$$

The first of these results, i.e. that for $f_{0,0}$, gives the frequency in the central cell of a group of nine ordinates, 3 × 3, in the scheme. The other expressions may be useful, where it is impossible to use a central cell, for example, towards the boundary in crateroid surfaces.

Let us put $\lambda_{0,1} = z_{0,1} + z_{0,-1} + z_{1,0} + z_{-1,0}$

and $\lambda_{1,1} = z_{1,1} + z_{1,-1} + z_{-1,1} + z_{-1,-1}$;

further we may write

$$\beta_{1,1} = f_{1,1} + f_{1,-1} + f_{-1,1} + f_{-1,-1}, \text{ and } \beta_{0,1} = f_{0,1} + f_{0,-1} + f_{1,0} + f_{-1,0}.$$

The first of the above equations gives us:

$$576 f_{0,0} = 484 z_{0,0} + 22\lambda_{0,1} + \lambda_{1,1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\kappa).$$

The sum of the next four provides:

$$576\beta_{1,1} = 16 z_{0,0} - 104\lambda_{0,1} + 676\lambda_{1,1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (\lambda),$$

while the sum of the last four leads to

$$576\beta_{0,1} = -176 z_{0,0} + 568\lambda_{0,1} + 52\lambda_{1,1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mu).$$

Multiply (κ) by 52 and subtract (μ) from it, we find:

$$29952 f_{0,0} - 576\beta_{0,1} = 25344 z_{0,0} + 576\lambda_{0,1} \quad \ldots\ldots\ldots\ldots\ldots\ldots (\nu).$$

Multiply (μ) by 13 and subtract (λ) from it, we have:

$$7488\beta_{0,1} - 576\beta_{1,1} = -2304 z_{0,0} + 7488\lambda_{0,1} \;. \quad \ldots\ldots\ldots\ldots\ldots (\xi).$$

Multiply (ν) by 13 and subtract (ξ) from it, there results:

$$389376 f_{0,0} - 14976\beta_{0,1} + 576\beta_{1,1} = 331776 z_{0,0} \quad \ldots\ldots\ldots\ldots\ldots (o).$$

Substituting for $\beta_{0,1}$ and $\beta_{1,1}$ we obtain after dividing by the coefficient of $z_{0,0}$:

$$z_{0,0} = \tfrac{1}{576}\{676 f_{0,0} - 26(f_{0,1} + f_{0,-1} + f_{1,0} + f_{-1,0}) + f_{1,1} + f_{1,-1} + f_{-1,1} + f_{-1,-1}\} \; (\pi).$$

This is the desired expression for the central ordinate in terms of the frequencies of nine adjacent cells. We suppose all cell bases equal; if they are not 1 × 1 squares, but rectangles $a \times b$, then the coefficient $\tfrac{1}{576}$ must be replaced by $\dfrac{1}{576ab}$.

We deduce from (κ) and (λ)

$$\lambda_{0,1} = \tfrac{1}{576}(208 f_{0,0} + 568\beta_{0,1} - 44\beta_{1,1}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (\rho),$$

$$\lambda_{1,1} = \tfrac{1}{576}(16 f_{0,0} + 88\beta_{0,1} + 484\beta_{1,1}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (\sigma).$$

By aid of these equations, I have solved the remaining equations for the ordinates in terms of the volumes. These are of value in finding ordinates at the boundaries of correlation tables. They are:

$$z_{0,1} = \tfrac{1}{576}\{52 f_{0,0} + 598 f_{0,1} - 26 f_{0,-1} - 2(f_{1,0} + f_{-1,0}) + f_{1,1} + f_{-1,1} - 23(f_{-1,-1} + f_{1,-1})\},$$
$$z_{1,0} = \tfrac{1}{576}\{52 f_{0,0} + 598 f_{1,0} - 26 f_{-1,0} - 2(f_{0,1} + f_{0,-1}) + f_{1,1} + f_{1,-1} - 23(f_{-1,-1} + f_{-1,1})\},$$
$$z_{0,-1} = \tfrac{1}{576}\{52 f_{0,0} + 598 f_{0,-1} - 26 f_{0,1} - 2(f_{1,0} + f_{-1,0}) + f_{1,-1} + f_{-1,-1} - 23(f_{-1,1} + f_{1,1})\},$$
$$z_{-1,0} = \tfrac{1}{576}\{52 f_{0,0} + 598 f_{-1,0} - 26 f_{1,0} - 2(f_{0,-1} + f_{0,1}) + f_{-1,1} + f_{-1,-1} - 23(f_{1,-1} + f_{1,1})\},$$
$$\ldots\ldots (\tau),$$

and:

$$z_{1,1} = \tfrac{1}{576}\{4 f_{0,0} + 276 f_{1,1} + 254 f_{-1,-1} - 23(f_{1,-1} + f_{-1,1}) + 46(f_{0,1} + f_{1,0}) - 2(f_{0,-1} + f_{-1,0})\},$$
$$z_{1,-1} = \tfrac{1}{576}\{4 f_{0,0} + 276 f_{1,-1} + 254 f_{-1,1} - 23(f_{1,1} + f_{-1,-1}) + 46(f_{1,0} + f_{0,-1}) - 2(f_{-1,0} + f_{0,1})\},$$
$$z_{-1,1} = \tfrac{1}{576}\{4 f_{0,0} + 276 f_{-1,1} + 254 f_{1,-1} - 23(f_{-1,-1} + f_{1,1}) + 46(f_{-1,0} + f_{0,1}) - 2(f_{0,-1} + f_{1,0})\},$$
$$z_{-1,-1} = \tfrac{1}{576}\{4 f_{0,0} + 276 f_{-1,-1} + 254 f_{1,1} - 23(f_{1,-1} + f_{-1,1}) + 46(f_{0,-1} + f_{-1,0}) - 2(f_{1,0} + f_{0,1})\},$$
$$\ldots\ldots (\upsilon).$$

The above are useful, but of course only approximative, methods of passing from ordinates to volumes or from volumes to ordinates in the case of frequency surfaces.

ON A SKEW CORRELATION SURFACE.

By E. C. RHODES, M.Sc.*

Professor Edgeworth has considered as a skew correlation surface of interest the surface

$$z = Z - \tfrac{1}{6}(q_{30}Z_{30} + 3q_{21}Z_{21} + 3q_{12}Z_{12} + q_{03}Z_{03}),$$

where

$$Z = \frac{1}{2\pi\sqrt{1-r^2}}\, e^{-\frac{1}{2}\frac{1}{1-r^2}(x^2-2rxy+y^2)},$$

expressing the variables x, y in terms of their standard deviations, where

$$Z_{mn} = \frac{d^{m+n}Z}{dx^m dy^n},$$

and where the q's are other constants involved in the surface, which can be shown to be moment coefficients.

From the equation

$$\log Z = -\tfrac{1}{2}C(x^2 - 2rxy + y^2) - \log\left(\frac{2\pi}{\sqrt{C}}\right),$$

writing

$$C = \frac{1}{1-r^2}$$

we get

$$\begin{aligned}
Z_{10} &= -CZ(x-ry), & Z_{01} &= -CZ(y-rx),\\
Z_{20} &= -CZ_{10}(x-ry) - CZ & Z_{02} &= -CZ_{01}(y-rx) - CZ\\
&= -CZ + C^2Z(x-ry)^2, & &= -CZ + C^2Z(y-rx)^2,\\
Z_{30} &= 3C^2Z(x-ry) - C^3Z(x-ry)^3, & Z_{03} &= 3C^2Z(y-rx) - C^3Z(y-rx)^3,
\end{aligned}$$

$$Z_{11} = -CZ_{01}(x-ry) + rCZ = rCZ + C^2Z(x-ry)(y-rx),$$

$$Z_{21} = C^3Z(y-rx) - 2rC^2Z(x-ry) - C^3Z(x-ry)^2(y-rx),$$

$$Z_{12} = C^3Z(x-ry) - 2rC^2Z(y-rx) - C^3Z(y-rx)^2(x-ry).$$

1. *Moments of the surface.* We will assume that x and y may vary from $-\infty$ to $+\infty$, and all integrations will be taken between these limits. See Note, p. 326.

(i) We have:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[Z - \tfrac{1}{6}(q_{30}Z_{30} + 3q_{21}Z_{21} + 3q_{12}Z_{12} + q_{03}Z_{03})]\,dx\,dy.$$

---

* Received June 4, 1925.

Now
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z\,dx\,dy = 1.$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{30}\,dx\,dy = \int_{-\infty}^{\infty} [Z_{20}]_{-\infty}^{\infty}\, dy = 0,$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{21}\,dx\,dy = [Z_{10}]_{-\infty}^{\infty} = 0,$$
because Z vanishes for extreme x and y, and similarly for the other integrals.

Consequently the volume of the surface is unity.

(ii) Consider $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zx\,dx\,dy$. We know $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Zx\,dx\,dy = 0$.
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{30}x\,dx\,dy = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{20}\,dx\,dy = -\int_{-\infty}^{\infty} [Z_{10}]_{-\infty}^{\infty}\, dy = 0,$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{21}x\,dx\,dy = \int_{-\infty}^{\infty} [Z_{20}]_{-\infty}^{\infty}\, x\,dx = 0,$$
for the same reasons as before. So with the other integrals. So also with
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zy\,dx\,dy.$$

Consequently the mean of the surface is the origin.

(iii) Consider $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zx^2\,dx\,dy$. We know $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Zx^2\,dx\,dy = 1$.
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{30}x^2\,dx\,dy = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{20}2x\,dx\,dy = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{10}\,dx\,dy$$
$$= 2\int_{-\infty}^{\infty} [Z]_{-\infty}^{\infty}\, dy = 0,$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{21}x^2\,dx\,dy = \int_{-\infty}^{\infty} [Z_{20}]_{-\infty}^{\infty}\, x^2\,dx = 0.$$

Similarly with the other integrals. So the standard deviation of the x-variable is unity. Similarly with the y-variable.

(iv) Consider $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zxy\,dx\,dy$. We know $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Zxy\,dx\,dy = r$.
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{30}xy\,dx\,dy = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{20}y\,dx\,dy = -\int_{-\infty}^{\infty} [Z_{10}]_{-\infty}^{\infty}\, y\,dy = 0,$$
$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{21}xy\,dx\,dy = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{11}y\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{10}\,dx\,dy$$
$$= \int_{-\infty}^{\infty} [Z]_{-\infty}^{\infty}\, dy = 0.$$

Similarly with the other integrals.

Consequently r is the correlation coefficient between the variables.

(v) Consider $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zx^3\,dx\,dy$. We know $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Zx^3\,dx\,dy = 0$.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{30}x^3\,dx\,dy = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{20}3x^2\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Z_{10}6x\,dx\,dy$$

$$= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 6Z\,dx\,dy = -6.$$

As before the other integrals are zero.

Consequently the third moment coefficient of the x-variable is q_{30}.

We find in exactly the same way that q_{03} is the third moment coefficient of the y-variable, and that q_{21}, q_{12} are the product moment coefficients obtained from Σx^2y, Σxy^2 respectively. As mentioned above the q's are the higher moment coefficients of the surface.

2. *The array.*

(i) Consider $\int_{-\infty}^{\infty} z\,dx$, keeping y constant. Call this z_y; the sum of the array.

We know $$\int_{-\infty}^{\infty} Z\,dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} = v\ (\text{say}).$$

$$\int_{-\infty}^{\infty} Z_{30}\,dx = [Z_{20}]_{-\infty}^{\infty} = 0, \text{ so also with } Z_{21},\ Z_{12},$$

$$\int_{-\infty}^{\infty} Z_{03}\,dx = \frac{d^3}{dy^3}\int_{-\infty}^{\infty} Z\,dx = \frac{d^3}{dy^3}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\right) = v_3\ (\text{say}).$$

Thus $$z_y = v - \tfrac{1}{6} q_{03} v_3 = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\left[1 - \frac{q_{03}}{2}\left(y - \frac{y^3}{3}\right)\right].$$

Similarly $$z_x = \int_{-\infty}^{\infty} z\,dx = u - \tfrac{1}{6} q_{30} u_3,$$

where $$u = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \text{ and } u_3 = \frac{d^3u}{dx^3}.$$

Thus the marginal curves of the surface are of the type developed by Edgeworth and others.

(ii) Consider $\int_{-\infty}^{\infty} zx\,dx$, keeping y constant. Call this $z_y\bar{x}_y$.

We know $$\int_{-\infty}^{\infty} Zx\,dx = vry.$$

$$\int_{-\infty}^{\infty} Z_{30}x\,dx = [Z_{20}x]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} Z_{20}\,dx = 0,$$

$$\int_{-\infty}^{\infty} Z_{21}x\,dx = [Z_{11}x]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} Z_{11}\,dx = 0,$$

$$\int_{-\infty}^{\infty} Z_{12}x\,dx = [Z_{02}x]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} Z_{02}\,dx = -\frac{d^2}{dy^2}\int_{-\infty}^{\infty} Z\,dx = -v_2\ (\text{say}),$$

$$\int_{-\infty}^{\infty} Z_{03}x\,dx = \frac{d^3}{dy^3}\int_{-\infty}^{\infty} Zx\,dx = \frac{d^3}{dy^3}(vry) = v_3ry + 3v_2r.$$

So we have
$$z_y \bar{x}_y = vry + \tfrac{1}{2} q_{12} v_2 - \tfrac{1}{6} q_{03} (v_3 ry + 3v_2 r)$$
$$= ry\,(v - \tfrac{1}{6} q_{03} v_3) + \tfrac{1}{2} v_2 (q_{12} - rq_{03}),$$

therefore
$$\bar{x}_y = ry + \tfrac{1}{2} (q_{12} - rq_{03}) \frac{v_2}{v - \frac{1}{6} q_{03} v_3}$$
$$= ry - \tfrac{1}{2} (q_{12} - rq_{03}) \frac{1 - y^2}{1 - \frac{1}{2} q_{03} \left(y - \dfrac{y^3}{3}\right)},$$

which gives the equation to a regression line. The other is similarly given by
$$\bar{y}_x = rx - \tfrac{1}{2} (q_{21} - rq_{30}) \frac{1 - x^2}{1 - \frac{1}{2} q_{30} \left(x - \dfrac{x^3}{3}\right)}.$$

It will be seen that the regression lines, in general, are not straight, and the conditions of rectilinearity are $q_{12} = rq_{03}$ and $q_{21} = rq_{30}$. These conditions have been obtained before and are well known.

(iii) Consider $\int_{-\infty}^{\infty} zx^2 dx$, keeping y constant. Call this $z_y(\sigma_y^2 + \bar{x}_y^2)$.

We know
$$\int_{-\infty}^{\infty} Zx^2 dx = v\,[1 - r^2 + r^2 y^2].$$
$$\int_{-\infty}^{\infty} Z_{30} x^2 dx = -\int_{-\infty}^{\infty} Z_{20} 2x\,dx = \int_{-\infty}^{\infty} Z_{10} 2\,dx = 0,$$
$$\int_{-\infty}^{\infty} Z_{21} x^2 dx = -\int_{-\infty}^{\infty} Z_{11} 2x\,dx = \int_{-\infty}^{\infty} Z_{01} 2\,dx = 2 \frac{d}{dy}(v) = 2v_1,$$
$$\int_{-\infty}^{\infty} Z_{12} x^2 dx = -\int_{-\infty}^{\infty} Z_{02} 2x\,dx = -2 \frac{d^2}{dy^2} \int_{-\infty}^{\infty} Zx\,dx = -2 \frac{d^2}{dy^2}(vry),$$
$$\int_{-\infty}^{\infty} Z_{03} x^2 dx = \frac{d^3}{dy^3} \int_{-\infty}^{\infty} Zx^2 dx = \frac{d^3}{dy^3} [v\,(1 - r^2 + r^2 y^2)].$$

So we have
$$z_y(\sigma_y^2 + \bar{x}_y^2) = v\,(1 - r^2 + r^2 y^2) - \tfrac{1}{6} \{3q_{21} \times 2v_1 - 3q_{12} \times 2\,(v_2 ry + 2v_1 r)$$
$$+ q_{03} [v_3 (1 - r^2 + r^2 y^2) + 3v_2 \times 2r^2 y + 3v_1 \times 2r^2]\}$$
$$= v\,(1 - r^2 + r^2 y^2) - q_{21} v_1 + rq_{12} (v_2 y + 2v_1)$$
$$- \frac{q_{03}}{6} [v_3 (1 - r^2 + r^2 y^2) + 6v_2 r^2 y + 6v_1 r^2]$$
$$= (v - \tfrac{1}{6} q_{03} v_3)(1 - r^2 + r^2 y^2) - v_1 (q_{21} - 2rq_{12} + r^2 q_{03})$$
$$+ v_2\, ry\,(q_{12} - rq_{03}),$$

therefore
$$\sigma_y^2 + \bar{x}_y^2 = 1 - r^2 + r^2 y^2 - \frac{(q_{21} - 2rq_{12} + r^2 q_{03})\, v_1}{v - \frac{1}{6} q_{03} v_3} + \frac{v_2 ry\,(q_{12} - rq_{03})}{v - \frac{1}{6} q_{03} v_3}.$$

But
$$\bar{x}_y^2 = r^2 y^2 + \frac{v_2 ry\,(q_{12} - rq_{03})}{v - \frac{1}{6} q_{03} v_3} + \tfrac{1}{4} \frac{(q_{12} - rq_{03})^2 v_2^2}{(v - \frac{1}{6} q_{03} v_3)^2},$$

therefore
$$\sigma_y^2 = 1 - r^2 - \frac{(q_{21} - 2rq_{12} + r^2 q_{03})\, v_1}{v - \frac{1}{6} q_{03} v_3} - \frac{1}{4}\frac{(q_{12} - rq_{03})^2\, v_2^2}{(v - \frac{1}{6} q_{03} v_3)^2}$$
$$= 1 - r^2 + \frac{(q_{21} - 2rq_{12} + r^2 q_{03})\, y}{1 - \frac{1}{2} q_{03}\left(y - \frac{y^3}{3}\right)} - \frac{1}{4}\frac{(q_{12} - rq_{03})^2 (1 - y^2)^2}{\left[1 - \frac{1}{2} q_{03}\left(y - \frac{y^3}{3}\right)\right]^2},$$
which gives the standard deviation of the array.

From this last we may obtain the correlation ratio, thus:

$$1 - \eta_y^2 = \int_{-\infty}^{\infty} \sigma_y^2 z_y\, dy = (1 - r^2)\int_{-\infty}^{\infty} z_y\, dy + (q_{21} - 2rq_{12} + r^2 q_{03})\int_{-\infty}^{\infty} v_1\, dy$$
$$- \frac{1}{4}(q_{12} - rq_{03})^2 \int_{-\infty}^{\infty} \frac{v_2^2}{v - \frac{1}{6} q_{03} v_3}\, dy,$$

therefore
$$\eta_y^2 = r^2 + \frac{1}{4}(q_{12} - rq_{03})^2 \int_{-\infty}^{\infty} \frac{v_2^2}{v - \frac{1}{6} q_{03} v_3}\, dy.$$

[η_y is the correlation ratio obtained by considering the standard deviation of y-arrays.]

The integral in this expression simplifies when $q_{03} = 0$.

Then we have
$$\int_{-\infty}^{\infty} (1 - y^2)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}\, dy = 2.$$

Thus we have as a simple case, when the marginal curve is a normal curve,
$$\eta_y^2 = r^2 + \frac{1}{2} q_{12}^2.$$

In the extreme case when q_{03} and $q_{30} = 0$, and both marginal curves are Gaussian, we have the regression lines
$$\bar{x}_y = ry - \frac{1}{2} q_{12}(1 - y^2),$$
$$\bar{y}_x = rx - \frac{1}{2} q_{21}(1 - x^2),$$
i.e. parabolas, and
$$\eta_y^2 = r^2 + \frac{1}{2} q_{12}^2,$$
$$\eta_x^2 = r^2 + \frac{1}{2} q_{21}^2.$$

The standard deviations of the arrays are given by
$$\sigma_y^2 = 1 - r^2 + (q_{21} - 2rq_{12})\, y - \frac{1}{4} q_{12}^2 (1 - y^2)^2,$$
$$\sigma_x^2 = 1 - r^2 + (q_{12} - 2rq_{21})\, x - \frac{1}{4} q_{21}^2 (1 - x^2)^2.$$

This is a case of a skew surface with normal curves for the margins and parabolic regression.

3. Let us consider the problem of finding the relationship between two variates when the individuals are arranged in order of merit, the problem of ranks and grades. Professor Karl Pearson has developed the method in *Drapers' Company Research Memoir*, Biometric Series, No. IV.

The essential parts of the argument only will be reproduced here. N individuals are distributed in a frequency group according to the extent to which they possess a variable quantity x. Let the frequency curve be $z = Nf(x)$, which ranges from $x = -a_1$ to $x = a_2$. Then the grade of the individual of character x

is $g(x) = \int_x^{a_2} Nf(x)\,dx$, and $\frac{dg}{dx} = -Nf(x)$. Mean grade is $\frac{N}{2} = \bar{g}$. $\sigma_g{}^2 = \frac{N^2}{12}$ gives the standard deviation of the grades.

Now suppose that two correlated variables are considered, and let the surface be $z = NF(x, y)$; the marginal curves being $z = Nf(x)$, $z = N\phi(y)$. To get the correlation between grades we require the integral

$$I = \iint (g_1 - \bar{g}_1)(g_2 - \bar{g}_2)\,NF(xy)\,dx\,dy,$$

the integration being taken over the whole range of x, y, where g_1, g_2 represent the grades of the individual which possesses the two characters to the amount x, y respectively, when the individuals are arranged in order first with respect to the one character x, then with respect to the other character y. We have

$$\frac{dg_1}{dx} = -Nf(x), \quad \frac{dg_2}{dy} = -N\phi(y).$$

Differentiating I with regard to r,

$$\frac{dI}{dr} \quad \iint (g_1 - \bar{g}_1)(g_2 - \bar{g}_2)\,N\frac{dF}{dr}\,dx\,dy.$$

Now with certain surfaces $\qquad \frac{dF}{dr} = \frac{d^2F}{dx\,dy}$;

assuming that our surface is such an one,

$$\begin{aligned}\frac{dI}{dr} &= \iint (g_1 - \bar{g}_1)(g_2 - \bar{g}_2)\,N\frac{d^2F}{dx\,dy}\,dx\,dy \\ &= \iint N\,F\frac{dg_1}{dx}\,\frac{dg_2}{dy}\,dx\,dy = N^3\iint F f(x)\,\phi(y)\,dx\,dy.\end{aligned}$$

Now we will show that the surface considered in the first part of the paper satisfies this condition $\frac{dF}{dr} = \frac{d^2F}{dx\,dy}$. We are still supposing x and y are measured in terms of their standard deviations.

The surface we are considering is

$$z = Z - \tfrac{1}{6}(q_{30}Z_{30} + 3q_{21}Z_{21} + 3q_{12}Z_{12} + q_{03}Z_{03}).$$

And since $\qquad \frac{dZ}{dr} = \frac{d^2Z}{dx\,dy}$,

also $\qquad \frac{d}{dr}\left(\frac{d^{m+n}Z}{dx^m dy^n}\right) = \frac{d^{m+n+2}Z}{dx^{m+1}dy^{n+1}}$,

so for our surface $\qquad \frac{dz}{dr} = \frac{d^2z}{dx\,dy}$.

Consequently we can use the method of Professor Pearson, when we assume the frequency surface to be that considered here, instead of the normal surface which he considers. So we shall write

$$\frac{dI}{dr} = N^3\iint [Z - \tfrac{1}{6}(q_{30}Z_{30} + 3q_{21}Z_{21} + 3q_{12}Z_{12} + q_{03}Z_{03})]\,(u - \tfrac{1}{6}q_{30}u_3)(v - \tfrac{1}{6}q_{03}v_3)\,dx\,dy,$$

using the notation already adopted for the marginal curves. The expression to be integrated is

$$\begin{aligned} Zuv &- \tfrac{1}{6}\left(q_{30}uvZ_{30} + 3q_{21}uvZ_{21} + 3q_{12}uvZ_{12} + q_{03}uvZ_{03} + q_{30}u_3vZ + q_{03}uv_3Z\right) \\ &+ \tfrac{1}{36}\left[q_{30}{}^2u_3vZ_{30} + q_{03}{}^2uv_3Z_{03} + q_{30}q_{03}\left(uv_3Z_{30} + u_3vZ_{03} + u_3v_3Z\right)\right. \\ &\qquad \left. + 3q_{21}q_{03}uv_3Z_{21} + 3q_{12}q_{03}uv_3Z_{12} + 3q_{21}q_{30}u_3vZ_{21} + 3q_{12}q_{30}u_3vZ_{12}\right] \\ &- \tfrac{1}{216}\,q_{30}q_{03}u_3v_3\left(q_{30}Z_{30} + 3q_{21}Z_{21} + 3q_{12}Z_{12} + q_{03}Z_{03}\right) \quad \ldots\ldots\ldots\ldots(1). \end{aligned}$$

Now
$$Zuv = \frac{\sqrt{C}}{(2\pi)^2}\,e^{-\frac{1}{2}\left[x^2+y^2+C\left(x^2+y^2-2rxy\right)\right]}$$
$$= \frac{\sqrt{C}}{(2\pi)^2}\,e^{-\frac{1}{2}(C+1)\left[x^2+y^2-2\frac{rC}{C+1}xy\right]},$$

which we can write
$$\frac{\sqrt{C}\,s^2\sqrt{1-R^2}}{2\pi}\cdot\frac{1}{2\pi\sqrt{1-R^2}\,.\,s^2}\,e^{-\frac{1}{2}\frac{1}{1-R^2}\left(\frac{x^2}{s^2}+\frac{y^2}{s^2}-2\frac{Rxy}{s^2}\right)}$$

where
$$(1-R^2)\,s^2 = \frac{1}{C+1} \text{ and } \frac{R}{s^2(1-R^2)} = rC.$$

Thus
$$R = \frac{rC}{C+1} = \frac{r}{2-r^2},$$

and
$$s^2 = \frac{2-r^2}{4-r^2},\quad 1-R^2 = \frac{(1-r^2)(4-r^2)}{(2-r^2)^2}.$$

Thus
$$Zuv = \frac{1}{2\pi\sqrt{4-r^2}}\cdot\frac{1}{2\pi s^2\sqrt{1-R^2}}\,e^{-\frac{1}{2}\frac{1}{1-R^2}\left(\frac{x^2}{s^2}+\frac{y^2}{s^2}-2\frac{Rxy}{s^2}\right)} \quad\ldots\ldots\ldots(2),$$

and as R varies between ± 1, when r varies between ± 1, it will be seen that the equation $z = Zuv$ is that of a normal frequency surface, the total volume (V) of which is $\frac{1}{2\pi\sqrt{4-r^2}}$, the correlation coefficient being R and the standard deviations of the marginal curves both s.

The integration of the various terms of the expression (1) will reduce to the finding of various product moment coefficients of the surface (2).

Consider first the terms only involving the q's to the first order

$$\iint uvZ_{30}\,dx\,dy = \int\left[uZ_{20}\right]v\,dy - \iint u_1vZ_{20}\,dx\,dy,$$

remembering that u is a function of x only, v a function of y only. This is equal to

$$-\int\left[u_1Z_{10}\right]v\,dy + \iint u_2vZ_{10}\,dx\,dy = \int\left[u_2Z\right]v\,dy - \iint u_3vZ\,dx\,dy,$$

where the terms in brackets [ ] at each stage are zero, owing to the exponential terms in the integrand.

Now $u_3 = u\,(3x - x^3)$, and the final integral is zero, because it is equivalent to the finding of the first and third moment coefficients of the margin of a normal surface.

Remembering that $\frac{u_1}{u}, \frac{u_3}{u}, \frac{u_5}{u} \dots$ are equal to odd functions of x, that $\frac{u_2}{u}, \frac{u_4}{u}$ are equal to even functions of x; similarly with $\frac{v_1}{v}, \frac{v_3}{v}, \dots, \frac{v_2}{v}, \frac{v_4}{v}, \dots$, and that

$$\iint u\ v\ Z_{21} dx\,dy = -\iint u_2 v_1 Z\,dx\,dy,$$

$$\iint u\ v\ Z_{03} dx\,dy = -\iint u\ v_3 Z\,dx\,dy,$$

$$\iint u_3 v_3 Z_{30} dx\,dy = -\iint u_6 v_3 Z\,dx\,dy,$$

$$\iint u_3 v_3 Z_{21} dx\,dy = -\iint u_5 v_4 Z\,dx\,dy,$$

and so on; also remembering that all product moments of a normal surface of the type $p_{2k+1,2l}$, $p_{2k,2l+1}$ are zero, e.g.

$$p_{10} = p_{01} = p_{30} = p_{03} = p_{21} = p_{12} = p_{32} = p_{23} = p_{43} = p_{34} = \text{etc.}$$

are all zero for a normal surface, we shall find it possible to evaluate the expression (1).

We realise that the integrals arising from the terms involving the q's in the first degree, and in the third degree in the expression (1), are all equivalent to moments of a normal surface of the form $p_{2k+1,2l}$ or $p_{2k,2l+1}$ and are zero. The only part of the expression (1) which contributes anything after Zuv is the term with coefficient $\frac{1}{36}$, involving the q's to second order products.

Concentrating now on these we have to write down the product moments of the normal surface, whose volume is V, correlation coefficient R and standard deviation of each margin s. Calling

$$\iint Zuvx^m y^n dx\,dy = Vp_{m.n} s^{m+n},$$

we have $p_{40} = p_{04} = 3$, $p_{60} = p_{06} = 15$, $p_{31} = p_{13} = 3R$, $p_{22} = 1 + 2R^2$, $p_{51} = p_{15} = 15R$, $p_{42} = p_{24} = 3 + 12R^2$, $p_{33} = 9R + 6R^3$.

$$\iint u_3 v Z_{30} dx\,dy = -\iint u_6 v Z\,dx\,dy = -V(15s^6 - 45s^4 + 45s^2 - 15),$$

since $$u_6 = u(x^6 - 15x^4 + 45x^2 - 15).$$

$$\iint uv_3 Z_{03} dx\,dy = -\iint uv_6 Z\,dx\,dy = -V(15s^6 - 45s^4 + 45s^2 - 15).$$

In the same way:

$$\begin{aligned}\iint uv_3 Z_{30} dx\,dy &= -\iint u_3 v_3 Z_0\,dx\,dy = \iint u_3 v Z_{03} dx\,dy\\ &= -\iint uvZ(3x - x^3)(3y - y^3)\,dx\,dy, \text{ since } u_3 = u(3x - x^3),\\ &= -V[9Rs^2 - 18Rs^4 + (9R + 6R^3)s^6],\end{aligned}$$

$$\iint uv_3 Z_{21} dx\,dy = -\iint u_2 v_4 Z\,dx\,dy = -\iint uvZ(x^2 - 1)(3 - 6y^2 + y^4)\,dx\,dy,$$

since $$v_4 = v(3 - 6y^2 + y^4).$$

This integral $= + V[3 - 9s^2 + 3s^4 + 6(1 + 2R^2)s^4 - (3 + 12R^2)s^6]$.

$$\iint uv_3 Z_{12}\,dx\,dy = -\iint u_1 v_5 Z\,dx\,dy = -\iint uv Zx(15y - 10y^3 + y^5)\,dx\,dy,$$

since $v_5 = -v(15y - 10y^3 + y^5)$.

This integral $= - V[15Rs^2 - 30Rs^4 + 15Rs^6]$.

$$\iint u_3 v Z_{21}\,dx\,dy = -\iint u_5 v_1 Z\,dx\,dy = - V[15Rs^2 - 30Rs^4 + 15Rs^6],$$

$$\iint u_3 v Z_{12}\,dx\,dy = -\iint u_4 v_2 Z\,dx\,dy = \text{same as} \iint uv_3 Z_{21}\,dx\,dy \text{ above.}$$

So we get

$$\begin{aligned}\frac{1}{N^3}\frac{dI}{dr} = V[1 &+ \tfrac{1}{36}(q_{30}{}^2 + q_{03}{}^2)(15 - 45s^2 + 45s^4 - 15s^6) \\ &\qquad - \frac{q_{30}q_{03}}{36}(9Rs^2 - 18Rs^4 + (9R + 6R^3)s^6) \\ &+ \tfrac{1}{12}(q_{21}q_{03} + q_{12}q_{30})(3 - 9s^2 + (9 + 12R^2)s^4 - (3 + 12R^2)s^6) \\ &- \tfrac{1}{12}(q_{12}q_{03} + q_{21}q_{30})(15Rs^2 - 30Rs^4 + 15Rs^6)].\end{aligned}$$

Examining each part of this in turn

$$15(1 - 3s^2 + 3s^4 - s^6) = 15(1 - s^2)^3 = \frac{120}{(4 - r^2)^3}.$$

$$\begin{aligned}3Rs^2(3 - 6s^2 + 3s^4 + 2R^2s^4) &= 3Rs^2[3(1 - s^2)^2 + 2R^2s^4] \\ &= \frac{3r}{4 - r^2}\left[3 \cdot \frac{4}{(4 - r^2)^2} + 2\frac{r^2}{(4 - r^2)^2}\right] = \frac{6r(6 + r^2)}{(4 - r^2)^3}.\end{aligned}$$

$$\begin{aligned}3(1 - 3s^2 + 3s^4 - s^6 + 4R^2s^4(1 - s^2)) &= 3[(1 - s^2)^3 + 4R^2s^4(1 - s^2)] \\ &= 3\left[\frac{8}{(4 - r^2)^3} + \frac{8r^2}{(4 - r^2)^3}\right] = 24\frac{1 + r^2}{(4 - r^2)^3}.\end{aligned}$$

$$15Rs^2(1 - 2s^2 + s^4) = \frac{60r}{(4 - r^2)^3}, \text{ for } 1 - s^2 = \frac{2}{4 - r^2},\ Rs^2 = \frac{r}{4 - r^2}.$$

Thus

$$\begin{aligned}\frac{1}{N^3}\frac{dI}{dr} &= V\left[1 + \frac{10(q_{30}{}^2 + q_{03}{}^2)}{3(4 - r^2)^3} - \frac{q_{30}q_{03}}{6}\frac{r(6 + r^2)}{(4 - r^2)^3}\right. \\ &\qquad\left. + 2(q_{21}q_{03} + q_{12}q_{30})\frac{1 + r^2}{(4 - r^2)^3} - 5(q_{12}q_{03} + q_{21}q_{30})\frac{r}{(4 - r^2)^3}\right] \\ &= V\left[1 + \frac{\lambda_1 - \lambda_2 r(6 + r^2) + \lambda_3(1 + r^2) - \lambda_4 r}{(4 - r^2)^3}\right] \text{ (say)} \\ &= V\left[1 + \frac{\lambda_1 + \lambda_2 r(4 - r^2 - 10) - \lambda_3(4 - r^2 - 5) - \lambda_4 r}{(4 - r^2)^3}\right] \\ &\ V\left|1 + \frac{\lambda_1 + 5\lambda_3}{(4 - r^2)^3} - \frac{\lambda_3}{(4 - r^2)^2} + \frac{\lambda_2 r}{(4 - r^2)^2} - \frac{(10\lambda_2 + \lambda_4)r}{(4 - r^2)^3}\right].\end{aligned}$$

So

$$2\pi\frac{dI}{N^3} = \frac{dr}{\sqrt{4 - r^2}}\left[1 + \frac{\lambda_1 + 5\lambda_3}{(4 - r^2)^3} - \frac{\lambda_3}{(4 - r^2)^2} + \frac{\lambda_2 r}{(4 - r^2)^2} - \frac{(10\lambda_2 + \lambda_4)r}{(4 - r^2)^3}\right].$$

Now if $r = 2\sin\theta$,

$$\int \frac{dr}{(4-r^2)^{\frac{5}{2}}} = \int \frac{2\cos\theta\, d\theta}{2^5 \cos^5\theta} = \frac{1}{2^4}\int \sec^4\theta\, d\theta = \frac{1}{2^4}\left(\tan\theta + \tfrac{1}{3}\tan^3\theta\right),$$

$$\int \frac{dr}{(4-r^2)^{\frac{7}{2}}} = \frac{1}{2^6}\left(\tan\theta + \tfrac{2}{3}\tan^3\theta + \tfrac{1}{5}\tan^5\theta\right),$$

$$\int \frac{r\,dr}{(4-r^2)^{\frac{5}{2}}} = \tfrac{1}{3}(4-r^2)^{-\frac{3}{2}} = \tfrac{1}{24}\sec^3\theta,$$

$$\int \frac{r\,dr}{(4-r^2)^{\frac{7}{2}}} = \tfrac{1}{5}(4-r^2)^{-\frac{5}{2}} = \tfrac{1}{160}\sec^5\theta.$$

Therefore

$$\begin{aligned}\frac{2\pi I}{N^3} + A \text{ (a constant)} &= \theta + \frac{\lambda_1 + 5\lambda_3}{64}\left(\tan\theta + \tfrac{2}{3}\tan^3\theta + \tfrac{1}{5}\tan^5\theta\right)\\ &\quad - \frac{\lambda_3}{16}\left(\tan\theta + \tfrac{1}{3}\tan^3\theta\right)\\ &\quad + \frac{\lambda_2}{24}\sec^3\theta - \frac{(10\lambda_2 + \lambda_4)}{160}\sec^5\theta\end{aligned}$$

$$\begin{aligned}&= \theta + \frac{\lambda_1}{64}\left(\tan\theta + \tfrac{2}{3}\tan^3\theta + \tfrac{1}{5}\tan^5\theta\right) + \frac{\lambda_2}{48}\left(2\sec^3\theta - 3\sec^5\theta\right)\\ &\quad + \frac{\lambda_3}{64}\tan\theta\sec^4\theta - \frac{\lambda_4}{160}\sec^5\theta.\end{aligned}$$

Now if ρ is the correlation between grades, $\rho = \dfrac{I}{N \cdot \dfrac{N^2}{12}}$.

Thus $\dfrac{I}{N^3} = \dfrac{\rho}{12}$, and when there is no correlation, ρ, r, θ vanish together, accordingly

$$A = -\frac{\lambda_2}{48} - \frac{\lambda_4}{160}.$$

Hence

$$\begin{aligned}\frac{\pi\rho}{6} &= \theta + \tfrac{5}{96}\left(q_{30}{}^2 + q_{03}{}^2\right)\left(\tan\theta + \tfrac{2}{3}\tan^3\theta + \tfrac{1}{5}\tan^5\theta\right)\\ &\quad + \frac{q_{30}q_{03}}{288}\left(1 + 2\sec^3\theta - 3\sec^5\theta\right)\\ &\quad + \frac{q_{21}q_{03} + q_{12}q_{30}}{32}\tan\theta\sec^4\theta - \frac{q_{12}q_{03} + q_{21}q_{30}}{32}\left(\sec^5\theta - 1\right) \quad \ldots\ldots\ldots\ldots(3).\end{aligned}$$

We may note that if $q_{30} = q_{03} = 0$, we get the same result as in the case of the normal surface, $\dfrac{\pi\rho}{6} = \theta$.

Thus $r = 2\sin\dfrac{\pi\rho}{6}$, in the case of a skew surface the margins of which are normal curves, and where the regression lines are parabolas. It must be remembered that the true measure of the relationship in this case between the variables is given by η.

If we suppose that the regression is nearly linear both ways we shall have $q_{21} = rq_{30}$ approximately and $q_{12} = rq_{03}$ approximately. If we substitute in (3) for these values we shall get

$$\frac{\pi\rho}{6} = \theta + \tfrac{5}{96}(q_{30}^2 + q_{03}^2)(\tan\theta + \tfrac{2}{3}\tan^3\theta + \tfrac{1}{5}\tan^5\theta)$$

$$+ \frac{q_{30}q_{03}}{288}(1 + 2\sec^3\theta - 3\sec^5\theta)$$

$$+ \frac{q_{30}q_{03}}{8}\tan^2\theta\sec^3\theta - \frac{q_{30}^2 + q_{03}^2}{16}\sin\theta(\sec^5\theta - 1)$$

$$= \theta + \frac{q_{30}^2 + q_{03}^2}{96}[5\tan\theta + \tfrac{10}{3}\tan^3\theta + \tan^5\theta - 6\tan\theta\sec^4\theta + 6\sin\theta]$$

$$+ \frac{q_{30}q_{03}}{288}[1 + 2\sec^3\theta - 3\sec^5\theta + 36\tan^2\theta\sec^3\theta]$$

$$= \theta + \frac{q_{30}^2 + q_{03}^2}{96}[-\tan\theta - \tfrac{26}{3}\tan^3\theta - 5\tan^5\theta + 6\sin\theta]$$

$$+ \frac{q_{30}q_{03}}{288}[33\sec^5\theta - 34\sec^3\theta + 1] \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4).$$

Let us consider a few values of r, with the corresponding θ's; if we denote by L the coefficient of $q_{30}^2 + q_{03}^2$, and by M the coefficient of $q_{30}q_{03}$, we have the following table:

TABLE I.

r	L	M
± 1	± ·0046	·0569
± ·9	± ·0095	·0394
± ·8	± ·0121	·0273
± ·7	± ·0129	·0187
± ·6	± ·0125	·0125
± ·5	± ·0113	·0081
± ·4	± ·0096	·0049
± ·3	± ·0075	·0026
± ·2	± ·0051	·0011
± ·1	± ·0026	·0003
0	·0000	·0000

Now in the method developed by Professor Pearson, enabling us to obtain the correlation coefficient from the correlation coefficient between ranks, which is used instead of the correlation coefficient between grades, we have $r = 2\sin\frac{\pi\rho}{6}$, on the assumption of normal correlation. Let us consider how much difference this assumption of normal correlation makes when the correlation surface is really skew. Equation (4) giving

$$\frac{\pi\rho}{6} = \theta + (q_{30}^2 + q_{03}^2)L + q_{30}q_{03}M \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5),$$

taken in conjunction with the preceding table, shows that $\frac{\pi\rho}{6} > \theta$ when q_{30} and q_{03} are both positive, if r and therefore θ are positive. $\frac{\pi\rho}{6}$ is greater numerically than θ when q_{30} and q_{03} are of opposite signs, if r and therefore θ are both negative. If r is positive and q_{30} and q_{03} are of opposite signs, $\frac{\pi\rho}{6}$ for certain values of q_{30} and q_{03} may be negative. The following table shows certain values of r calculated by the usual method when the surface is really skew, for $q_{30} = q_{03}$.

TABLE II.

Values of r calculated by Method of Ranks and Grades.

Values of $q_{30} = q_{03}$.

Real values of r.	*0·1*	*0·2*	*0·3*	*0·4*	*0·5*	*0·6*	*0·7*	*0·8*	*0·9*
1·0	1·001	1·005	1·010	1·018	1·029	1·041	1·055	1·072	1·091
·9	·901	·905	·910	·917	·926	·937	·951	·966	·984
·8	·801	·804	·809	·815	·823	·834	·846	·860	·876
·7	·701	·703	·707	·713	·720	·730	·741	·753	·767
·6	·601	·603	·606	·611	·617	·625	·634	·645	·658
·5	·500	·502	·505	·510	·515	·521	·529	·538	·548
·4	·400	·402	·404	·408	·412	·417	·423	·430	·438
·3	·300	·301	·303	·306	·309	·313	·317	·322	·328
·2	·200	·201	·202	·204	·206	·208	·211	·214	·218
·1	·100	·101	·101	·102	·103	·104	·105	·107	·109

For instance suppose the two variables are distributed according to a skew distribution where q_{30} and q_{03} are 0·5 and the real coefficient of correlation between them is 0·6. Then θ is obtained as the angle whose sine is 0·3.

The equation (5) above shows that the correlation by ranks would be given by

$$\frac{\pi\rho}{6} = \sin^{-1} 0\cdot3 + \frac{L}{2} + \frac{M}{4}.$$

Now if the correlation coefficient were calculated from this ρ in the usual manner we should find $2\sin\frac{\pi\rho}{6}$, which would be in this case ·617. Thus the error made in assuming normal correlation where the correlation is skew is in this case in excess by about 3 %. A glance at Table II shows that for the cases considered the errors (all in excess) are never more than 10 %. Equation (5) shows that for a given q_{30} (or q_{03}) the difference between $\frac{\pi\rho}{6}$ and θ is greatest when $q_{03} = q_{30}$ (or $q_{30} = q_{03}$), so we may say that the value of r as found by the usual formula $\left(r = 2\sin\frac{\pi\rho}{6}\right)$ is never more than 10 % out, so long as the third moment coefficients (q_3's) of the individual variable distributions are not greater than 0·9. Professor Edgeworth calls slightly abnormal frequency distributions for which q_3 is less than

0·5, and moderately abnormal distributions for which q_3 lies between 0·5 and 0·85 roughly. So we may say that when the skew distributions of the correlated variables are slightly abnormal, the error made in the calculation of **r** by the ranks correlation method assuming normal distributions is not more than 3 °/₀; and for moderately abnormal distributions the error is not more than 9 °/₀. Thus, when we consider that this method of finding the coefficient of correlation between two variables is generally used when the total frequency is not very large, and fairly large probable errors are involved, we realise that the error introduced, due to the distribution of frequency being slightly or moderately skew, is not very large, when the probable errors involved are taken into account.

Note on integration from $-\infty$ *to* $+\infty$. So long as the q's in the equation of the surface are small, the amount of negative frequency for values of x, y beyond the bounding curve given by

$$Z = \tfrac{1}{6}(q_{30}Z_{30} + 3q_{21}Z_{21} + 3q_{12}Z_{12} + q_{03}Z_{03})$$

is negligible, because the limits of this curve are distant more than three times the standard deviation from the mean. See a paper by the author in the *Statistical Journal*, Vol. LXXXVIII. Part IV, where this matter is discussed in the case of a moderately abnormal distribution.

ON THE EXCESS MORTALITY OF MALES IN THE FIRST YEAR OF LIFE.

BY MAJOR GREENWOOD AND E. M. NEWBOLD.

PROFESSOR F. LENZ of Munich contributes to the *Arch. f. Hygiene* (Bd. XCIII. S. 126—150) a paper entitled "Die Uebersterblichkeit der Knaben im Lichte der Erblichkeitslehre," which suggests some interesting reflections.

At the outset, Lenz remarks that the Mendelian inheritance of sex is now established ("kann heute als sichergestellt gelten"). Hence, he thinks, we can explain the fact, or alleged fact, that recessive pathological hereditary characters are only found in males. From these hypotheses—the correctness of which we do not intend to discuss—he infers that a larger fraction of total male mortality in infancy is selective than of female mortality. If this be so, it would follow, he suggests, that when a non-selective factor, or relatively non-selective factor, such as a prejudicial general environmental change, e.g. a hot summer or an outbreak of an epidemic, heightens the whole of the mortality of the first year of life, the *relative* excess of male mortality should be reduced. Lenz's method of testing this proposition is to correlate* the ratio of the rate of mortality on males to that on females with the rate of mortality on both sexes in 11 series of annual rates of mortality for different countries (in most cases the number of years included is very small, 8 or 10, in only one does it exceed 40) and he finds that in the great majority the value of r is large and negative.

It will be interesting to examine Lenz's argument from a biometric point of view. If x is the rate of male mortality (Lenz uses, and we shall use, the conventional measure, ratio of deaths under 1 to births in the year) and y that of female mortality, what Lenz is correlating is x/y and $\frac{Ax + By}{A + B}$. But in his series A and B are approximately constant and each is nearly equal to 0·5, so that approximately we have the correlation of x/y with $x+y$. To a first approximation this is equal to

---

* In addition to the ordinary product moment r, Lenz uses a coefficient invented by himself which he proposes to name the "deutscher Korrelations Index, k." It appears from the definition and worked example of this coefficient on p. 147, that if in a given set of data, there is no instance in which a value of one variable greater than its mean is associated with a value of the other variable less than its mean, the index must be equal to $+1$ or if every deviation in excess of one mean is associated with a deviation in defect of the other mean, the index will be -1, whatever the magnitude of the corresponding deviations. But on p. 132 there is a set of 8 observations (France) giving a product moment coefficient of $-·86$ and such that every positive deviation of one variable is associated with a negative deviation of the other variable. But the German index is stated to be not $-1·0$ but $-0·64$. We are therefore not sure that we know what the German index really is, and cannot therefore say what merits, if any, it possesses.

$$\frac{\dfrac{\sigma_x^2}{\bar{x}} - \dfrac{\sigma_y^2}{\bar{y}} + r_{xy}\,\sigma_x\sigma_y\left(\dfrac{1}{\bar{x}} - \dfrac{1}{\bar{y}}\right)}{\sqrt{\dfrac{\sigma_x^2}{\bar{x}^2} + \dfrac{\sigma_y^2}{\bar{y}^2} - \dfrac{2r_{xy}\,\sigma_x\sigma_y}{\bar{x}\bar{y}}}\,\sqrt{\sigma_x^2+\sigma_y^2+2r_{xy}\sigma_x\sigma_y}} \quad\text{.........}\ (1).$$

The last term of the numerator is always negative because r_{xy} is positive and $\bar{x}$ is greater than $\bar{y}$. So that if $\dfrac{\sigma_x^2}{\bar{x}} - \dfrac{\sigma_y^2}{\bar{y}}$ is not positive and greater in absolute value than $r_{xy}\sigma_x\sigma_y\left(\dfrac{1}{\bar{x}} - \dfrac{1}{\bar{y}}\right)$ then $r_{\frac{x}{y}(x+y)}$ must be negative. Obviously then if $\dfrac{\sigma_y^2}{\bar{y}} > \dfrac{\sigma_x^2}{\bar{x}}$ the value of $r_{\frac{x}{y}(x+y)}$ is always negative.

Let us examine the condition for a negative correlation more closely. Write

b = male births, d = male deaths,
b' = female births, d' = female deaths,

and assume a constant sex ratio, viz. $b = kb'$.

Let $$M = \frac{d+d'}{b+b'}\,;\ m = \frac{d}{b} \text{ and } m' = \frac{d'}{b'}\,.$$

Then $$M = \frac{km}{1+k} + \frac{m'}{1+k} \text{ and } \delta M = \frac{1}{1+k}(k\delta m + \delta m').$$

Put $z = \dfrac{m}{m'}$, so that, to a first approximation, $\dfrac{\delta z}{z} = \dfrac{\delta m}{m} - \dfrac{\delta m'}{m'}$

Summing $\delta M\delta z$, we have

$$r_{Mz}\,\sigma_M\sigma_z = \frac{1}{\bar{m}'^2(1+k)}\left[k\bar{m}'\,\sigma_m^2 - \bar{m}\sigma_{m'}^2 + r_{mm'}\,\sigma_m\sigma_{m'}\,(\bar{m}' - k\bar{m})\right].$$

Then writing $$v_m = \frac{\sigma_m}{\bar{m}},\ v_m' = \frac{\sigma_{m'}}{\bar{m}'},$$

the sign of r_{Mz} is the same as that of

$$k\bar{m}v_m^2 - \bar{m}'v_{m'}^2 + r_{mm'}\,v_m v_{m'}\,(\bar{m}' - k\bar{m}),$$

i.e. of $$k\bar{m} - \bar{m}'h^2 + r_{mm'}\,h\,(\bar{m}' - k\bar{m}),\ \text{where } h = \frac{v_m'}{v_m},$$

which is negative if the positive root of

$$\bar{m}'h^2 - hr_{mm}\,(\bar{m}' - k\bar{m}) - k\bar{m} = 0$$

is less than h,

i.e. if $$\frac{1}{2\bar{m}'}\left[r_{mm'}(\bar{m}' - k\bar{m}) + \sqrt{r_{mm'}^2(\bar{m}' - k\bar{m})^2 + 4k\bar{m}'\bar{m}}\right]$$

is less than h. For $r_{mm'} = 1$, the root is 1; thus, since $r_{mm'}$ is of the order of ·97, the root in general differs little from unity, and (since $k \nless 1$ and $m \nless m'$) usually, but not necessarily, exceeds unity.

So far as the short series given by Lenz are concerned, (1) seems adequate. Thus he gives the following figures for Hungary 1906—15.

	Male Mortality per 1000	Female Mortality per 1000	Both sexes per 1000	Per cent. excess of Male/Female
1906	22·0	18·8	20·5 (20·4)	17·0
1907	22·5	19·1	20·8 (20·8)	17·8
1908	21·5	18·2	19·9 (19·9)	18·1
1909	22·8	19·5	21·2 (21·2)	16·9
1910	20·9	17·9	19·4 (19·4)	16·8
1911	22·3	19·0	20·7 (20·7)	17·4
1912	20·0	17·1	18·6 (18·6)	17·0
1913	21·7	18·5	20·1 (20·1)	17·3
1914	21·0	17·9	19·5 (19·5)	17·3
1915	28·2	24·5	26·4 (26·4)	15·1

Save in the first case, the sex ratio is, to the number of places given, unity.

Lenz finds by direct calculation (we have verified this) that the correlation of the last two columns is $-\cdot 77$ and, if the data for 1915 are omitted, $+\cdot 16$ (according to us $+\cdot 17$).

From columns (2) and (3) we deduce

$$\begin{array}{rl}
 & \text{omitting 1915} \\
\bar{x} = 22\cdot 29, & 21\cdot 63, \\
\sigma_x = 2\cdot 12, & \cdot 838, \\
\bar{y} = 19\cdot 05, & 18\cdot 444, \\
\sigma_y = 1\cdot 94, & \cdot 703, \\
v_x = \cdot 095, & \cdot 0387, \\
v_y = \cdot 102, & \cdot 0381, \\
r_{xy} = \cdot 999, & \cdot 996.
\end{array}$$

Using (1) we have $r_{\frac{x}{y}(x+y)} = -\cdot 81$ and $+\cdot 18$ respectively, a sufficiently good approximation. Of course the data are too scanty to be of more than arithmetical interest.

The same negative correlation has been pointed out by the Registrar General for England and Wales in his Statistical Review for 1921 (p. 22 *et seq.*). In an interesting discussion he estimates the share in this increasing male excess which must result as a natural consequence of the greater male mortality in the earlier months of the first year of life to be a comparatively small one. He examines the facts in the light of the theory that a larger proportion of the males are essentially non-viable, a somewhat similar theory to that of Lenz; but as the excess mortality of males is less in the cases of prematurity and congenital malformation (causes from which a large proportion at least of deaths may be assumed to be inevitable) than for All Causes, he finds the facts difficult of explanation, and reaches no definite conclusion.

That the general variability of mortality in the first year of life is greater in females than in males in England and Wales sufficiently appears from the following examples. Taking first the 55 Registration Counties (experience 1901—10) we find:

$$\begin{aligned} \bar{x} &= 125{\cdot}2, \\ \bar{y} &= 100{\cdot}82, \\ \sigma_x &= 20{\cdot}46, \\ \sigma_y &= 17{\cdot}66, \\ v_x &= {\cdot}16, \\ v_y &= {\cdot}18. \end{aligned}$$

Taking 86 Registration Districts wherein the total births in 1901—10 were between 10,000 and 20,000:

$$\begin{aligned} \bar{x} &= 125{\cdot}05, \\ \bar{y} &= 99{\cdot}84, \\ \sigma_x &= 18{\cdot}17, \\ \sigma_y &= 15{\cdot}62, \\ v_x &= {\cdot}15, \\ v_y &= {\cdot}16. \end{aligned}$$

Taking (1) 91 Registration Districts with proportion of Male Births from ·506—·508, and (2) 84 Registration Districts with proportion of Male Births from ·508—·510, we have:

	(1)	(2)
$\bar{x} =$	$129{\cdot}637$,	$127{\cdot}179$,
$\bar{y} =$	$104{\cdot}089$,	$101{\cdot}048$,
$\sigma_x =$	$31{\cdot}999$,	$31{\cdot}556$,
$\sigma_y =$	$27{\cdot}807$,	$28{\cdot}410$,
$v_x =$	${\cdot}247$,	${\cdot}248$,
$v_y =$	${\cdot}267$,	${\cdot}281$.

In each case v_y exceeds v_x.

We propose now to examine the variabilities of Specific causes of death.

The groups chosen were (*a*) All Causes, (*b*) Measles, (*c*) Whooping Cough, (*d*) Congenital Debility, (*e*) Diphtheria and Croup, (*f*) Tuberculosis (all kinds), (*g*) Other Diseases of the Respiratory Organs, (*h*) Diarrhoeal Diseases.

The exact list of headings included in each of these groups is given in Table A of the Appendix. Owing to changes in classification, and varying degrees of accuracy of assignment to cause over a long period, some caution would probably have to be observed if we were comparing the absolute values of these rates at different periods, but as we are dealing with the relative male and female rates this consideration is not of much weight.

We have varied our unit of observation to get (using Tschouproff's terms) both a static and a dynamic variation. For the dynamic or secular we have taken the calendar year as unit, and traced the infant mortality of England and Wales, and of London, separately down to 1922; the date of beginning varies for some causes, but is usually 1847 for England and Wales, and 1845 for London. In some cases, where for one reason or another a break is indicated, we have tested our results by recalculating for shorter periods. For the static variation we have taken first a geographical variation, with the unit of observation a Registration County* of England and Wales (Urban and Rural separately), all for the period 1911—1923 inclusive, and secondly an occupational variation with the unit of observation the occupation† of the child's father. This last is solely for 1911 and is only available for All Causes. Before going into the results it is perhaps worth remembering that though we propose to compare male and female variation in mortality, we are not doing so in the same sense as when we compare male and female variation in individual characters such as height, weight, cephalic index, etc. In the one case the unit is an individual, and each individual can have any numerical value of the quality considered, in the other case the unit is a group of individuals—grouped either by time, place or occupation—and it is only in the case of the *group* that the quality can have a series of different numerical values. In other words to say that a set of groups of females is (absolutely or relatively) more variable than a set of groups of males under apparently the same conditions is to say that the same increment of adverse conditions to both, causes (absolutely or relatively) more additional female than male deaths.

Since we are dealing with the variation of a ratio, we will first see how much of the variation may be due to simple sampling, and how this part differs in the male and female mortality, though it is, of course, abundantly clear without any calculation at all, that our figures differ from those of a case of simple sampling in more than one respect.

Suppose q is the rate of mortality per person in any population and we take a series of samples of size n out of the population at random, the standard deviation of q will be $\sqrt{\frac{pq}{n}}$ where $p = 1 - q$, its Coefficient of Variation $100\sqrt{\frac{p}{nq}}$ and the quantity $\frac{\sigma_q^2}{q}$ $\left(\text{the } \frac{\sigma_x^2}{\bar{x}}\right.$ of equation (1), on which the sign of Lenz's correlation depends$\left.\right)$, will be reduced to $\frac{p}{n}$.

---

* The counties where the total births (Rural and Urban, male and female together) did not exceed 20,000 have been grouped as follows:

Anglesey and Carnarvon.
Brecknock and Cardigan.
Merioneth, Montgomery and Radnor.
Ely, Huntingdon, Peterborough and Rutland.
Westmorland and Cumberland.
Isle of Wight and Southampton.

This gave 52 Counties in each set.

† Only those occupations were included where the numbers of births of both male and female separately were over 1000.

The curves of these three variables for increasing values of q and constant value of n are shown in Figures 1, 2, and 3. The values of q with which we are dealing are all very small so that we are only concerned with the left-hand part of the graphs. For these it is clear that if simple sampling alone is in question, the standard deviation increases, while the coefficient of variation and the function $\frac{\sigma_q^2}{q}$ decrease, with increasing q for constant n. If, as in our cases, the numbers in the samples vary, we simply have to replace n by the Harmonic Mean of its various values. Hence if the numbers of births of boys and girls were really—instead of only approximately—equal, then since the male mortality is, for all cases except whooping cough, greater than the female, the male standard deviation would be greater but the male coefficient of variation, and also the function $\frac{\sigma_q^2}{q}$, *less* than the female. Hence the negative correlation observed by Lenz would occur, not only for all causes, but for each cause separately (except whooping cough, where it would be positive). As a matter of fact since the male births in reality always slightly exceed the female, n and q increase together, so that what variation there is in n tends (except for whooping cough, where the reverse is the case) in the case of the standard deviation to counteract its tendency to increase with q, and in the case of the other two functions shown in the figures to accentuate their tendency to decrease as q increases.

In all cases and causes, therefore,—where the male mortality exceeds the female and where simple sampling alone holds,—we expect, by virtue of equation (1), a negative correlation between the male/female and the total mortality.

As we have pointed out before, however, it is quite obvious that the simple sampling scheme cannot be expected to fit any one of our sets of series. Neither in the case of the calendar years, nor of the geographical districts, nor of the father's occupational groups, can the samples of our series be supposed to be drawn out of the same bag—we know that in many cases a real difference in the chance of infant death exists from year to year and from district to district. Such differences would tend to make our observed standard deviations greater than the theoretical*. Our series must also, however, depart from those of simple sampling in another direction, in any single draw of n balls from a bag—or to vary our analogy in any single throw of n dice the chance of throwing a given number or not, is not always the same from one die to another; in other words, the chance of death is not alike for everybody in one district, or one calendar year, or one occupational group; this is obviously true in any case and it is accentuated in the case of special causes whose incidence is largely seasonal or epidemic, and thus not spread evenly over the year of exposure. Infants for

---

* If s is the observed, and $\sigma = \sqrt{\frac{p_0 q_0}{n}}$ the theoretical standard deviation of q, then

$$s^2 = \frac{p_0 q_0}{n} + \frac{n-1}{n} \sigma_q^2,$$

where σ_q is the standard deviation of the different q's.

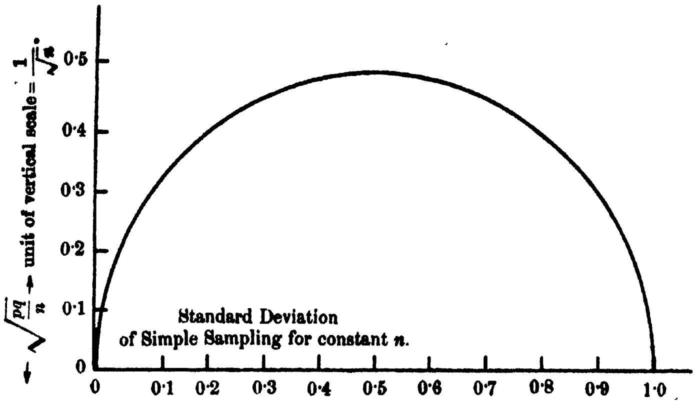


Fig. 1. Variation due to Simple Sampling.

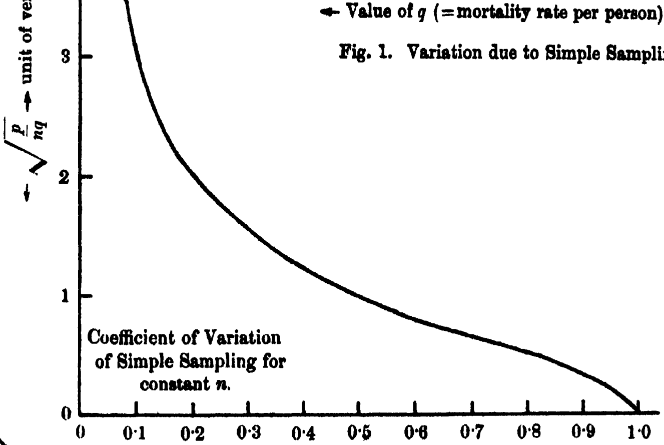


Fig. 2. Variation due to Simple Sampling.

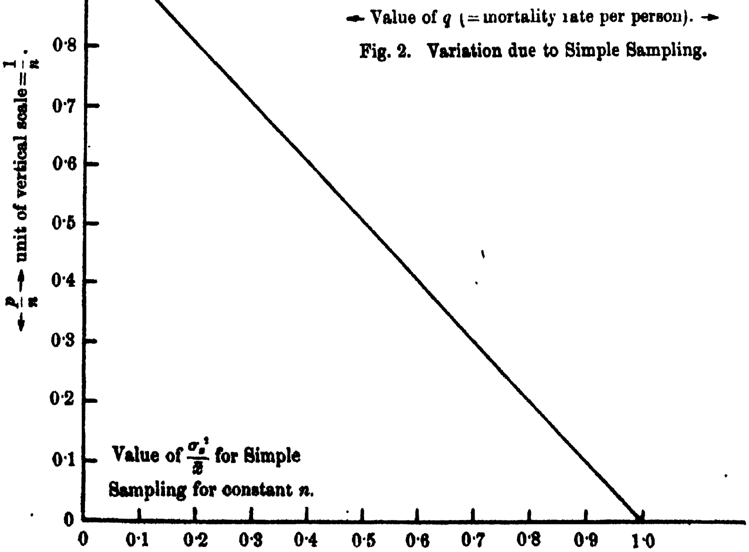


Fig. 3. Variation due to Simple Sampling.

instance, born in October, would have reached a more resistant age when the time came for the next summer's rise in infantile diarrhoea than those some few months later. This uneven weighting of individual dice tends to make the observed standard deviation less than the theoretical*.

Of these opposite tendencies our results show us that the first is in almost every case the stronger, in other words the effect of the differences between the groups is considerably greater than that of the differences between the chances of death of the individuals in a group, for in Table I, the observed standard deviation exceeds the theoretical for both male and female. (In the group Diphtheria and Croup alone, for the variation from county to county, though not in the secular variation, does the theoretical approximate to and in one case even slightly exceed the observed variation.) Hence on the whole the first tendency is clearly the more important to consider.

We have no *à priori* reason for supposing that this variation of the districts, years or occupations will follow the same sort of course with increasing q as does that part due to simple sampling, as we have no *à priori* knowledge of their distribution. From our empirical results in Table I, we see that as a matter of fact the nett observed standard deviations and coefficients of variation, do, the former always, the latter only on the whole and with several exceptions (e.g. especially in some of the groups for measles, diphtheria and croup, and respiratory diseases), behave in the same way as the simple sampling variations, i.e. in general the male standard deviation is greater than the female standard deviation and the female coefficient of variation more often greater than less than the male coefficient of variation (with the exception of whooping cough which is in both cases reversed). This general rule for the absolute and relative variation has been tested also for all causes, England and Wales, when the variation is measured by the average absolute difference between each pair of consecutive years, and the percentage of this difference on the rate for the first of the two years. For England and Wales (All Causes, 1847—1922) the figures are:

	Male	Female
Average absolute difference	8·81	8·08
Average percentage absolute difference ...	5·93	6·65

If we want to get a true idea of whether we are justified in attaching any sort of importance to the difference between male and female variations, we must first get rid of that part of it which may be due to random sampling and then examine the difference that remains.

The relative true variation—apart from chance—in the rates is shown in the last column of Table I, which gives the value of

$$\frac{100\sqrt{(\text{observed s. d. of rate})^2-(\text{simple sampling s. d. of rate})^2}}{\text{mean rate}},$$

---

[*] In this case if s is the observed, and $\sigma=\sqrt{\frac{p_0q_0}{n}}$ the theoretical standard deviation of q, then $s^2=\frac{p_0q_0}{n}-\frac{\sigma_q^2}{n}$, where σ_q is the standard deviation of the individual chances.

TABLE I (*a*). *Secular Variation in Infant Mortality by Causes and Sex.*

Group		Mean mortality per 1000	Standard Deviation		Coefficient of Variation		* (see below)
			Observed per 1000	Simple Sampling per 1000	Observed	Simple Sampling	
All Causes:							
England and Wales, 1839—1922	Male	153·85 ± 1·68	22·86 ± 1·19	·580	14·86 ± ·79	·377	14·85
	Female	126·19 ± 1·50	20·37 ± 1·06	·545	16·14 ± ·86	·432	16·14
London, 1839—1922	Male	157·90 ± 1·88	25·49 ± 1·33	1·606	16·14 ± ·86	1·017	16·11
	Female	132·50 ± 1·72	23·35 ± 1·22	1·523	17·62 ± ·95	1·149	17·59
Measles:							
England and Wales, 1847—1922	Male	2·66 ± ·057	·737 ± ·040	·081	27·70 ± 1·63	3·04	27·54
	Female	2·28 ± ·046	·600 ± ·033	·076	26·29 ± 1·53	3·35	26·10
England and Wales, 1847—1913	Male	2·74 ± ·053	·639 ± ·037	—	23·34 ± 1·43	—	—
	Female	2·35 ± ·042	·512 ± ·030	—	21·76 ± 1·33	—	—
London, 1845—1922	Male	3·48 ± ·087	1·137 ± ·061	·252	32·70 ± 1·95	7·25	44·71
	Female	2·99 ± ·076	·993 ± ·054	·231	33·22 ± 1·98	7·97	32·30
Whooping Cough:							
England and Wales, 1847—1922	Male	5·16 ± ·098	1·27 ± ·069	·112	24·54 ± 1·43	2·18	24·52
	Female	5·80 ± ·107	1·38 ± ·076	·122	23·80 ± 1·37	2·10	23·70
England and Wales, 1847—1913	Male	5·43 ± ·084	1·020 ± ·059	—	18·77 ± 1·132	—	—
	Female	6·10 ± ·091	1·098 ± ·064	—	17·99 ± 1·082	—	—
London, 1845—1922	Male	6·64 ± ·185	2·426 ± ·131	·348	36·56 ± 2·223	5·24	36·22
	Female	7·31 ± ·203	2·656 ± ·143	·372	36·34 ± 2·206	5·09	36·03
Diphtheria and Croup:							
England and Wales, 1855—1922	Male	·865 ± ·045	·545 ± ·032	·045	63·04 ± 4·885	5·23	62·79
	Female	·668 ± ·034	·415 ± ·024	·041	62·11 ± 4·781	6·07	61·83
London, 1859—1922	Male	1·094 ± ·049	·575 ± ·034	·136	52·57 ± 3·91	12·40	51·07
	Female	·851 ± ·034	·404 ± ·024	·120	47·40 ± 3·40	14·07	45·33
London, 1866—1922	Male	·961 ± ·039	·438 ± ·028	—	45·64 ± 3·43	—	—
	Female	·769 ± ·031	·341 ± ·022	—	44·40 ± 3·31	—	—
England and Wales, 1866—1922	Male	·686 ± ·031	·345 ± ·022	—	50·37 ± 3·91	—	—
	Female	·528 ± ·023	·254 ± ·016	—	48·13 ± 3·68	—	—
Tuberculosis (all kinds):							
England and Wales, 1847—1922	Male	8·465 ± ·241	3·115 ± ·170	·144	36·80 ± 2·27	1·70	36·82
	Female	6·768 ± ·195	2·524 ± ·138	·131	37·30 ± 2·31	1·94	37·18
London, 1845—1922	Male	11·68 ± ·367	4·802 ± ·259	·460	41·10 ± 2·57	3·94	40·91
	Female	9·44 ± ·377	4·932 ± ·266	·422	52·23 ± 3·51	4·47	52·03
Respiratory Diseases (other than Tuberculosis):							
England and Wales, 1847—1922	Male	26·626 ± ·327	4·225 ± ·231	·253	15·866 ± ·890	·949	15·86
	Female	20·755 ± ·258	3·335 ± ·183	·228	16·068 ± ·902	1·10	16·05
London, 1845—1922	Male	30·188 ± ·415	5·439 ± ·294	·733	18·018 ± ·415	2·43	17·86
	Female	24·130 ± ·328	4·288 ± ·232	·670	17·772 ± ·990	2·78	17·56
London, 1847—1913	Male	31·37 ± ·382	4·64 ± ·270	—	14·783 ± ·880	—	—
	Female	25·14 ± ·293	3·56 ± ·207	—	14·151 ± ·841	—	—
London, 1847—1922	Male	30·23 ± ·426	5·50 ± ·301	—	18·189 ± 1·028	—	—
	Female	24·21 ± ·334	4·32 ± ·236	—	17·824 ± 1·006	—	—
Diarrhoea:							
England and Wales, 1847—1922	Male	19·080 ± ·573	7·405 ± ·405	·215	38·812 ± 2·422	1·13	38·82
	Female	16·098 ± ·523	6·753 ± ·370	·202	41·953 ± 2·669	1·25	41·91
London, 1845—1922	Male	22·506 ± ·571	7·481 ± ·404	·635	33·242 ± 1·98	2·82	33·12
	Female	19·424 ± ·534	6·992 ± ·378	·603	35·995 ± 2·18	3·10	35·85
Congenital Debility:							
England and Wales, 1847—1922	Male	44·43 ± ·275	3·55 ± ·194	·323	7·99 ± ·440	·728	7·96
	Female	36·58 ± ·258	3·33 ± ·182	·301	9·11 ± ·502	·822	9·07
England and Wales, 1847—1913	Male	45·32 ± ·220	2·67 ± ·156	—	5·89 ± ·344	—	—
	Female	37·48 ± ·191	2·31 ± ·135	—	6·17 ± ·361	—	—
London, 1845—1922	Male	39·03 ± ·303	3·970 ± ·214	·830	10·17 ± ·555	2·13	9·95
	Female	32·36 ± ·250	3·275 ± ·177	·773	10·12 ± ·552	2·39	9·85
London, 1847—1922	Male	39·29 ± ·282	3·643 ± ·199	—	9·27 ± ·512	—	—
	Female	32·52 ± ·241	3·110 ± ·170	—	9·56 ± ·528	—	—
London, 1847—1913	Male	39·94 ± ·265	3·221 ± ·188	—	8·064 ± ·473	—	—
	Female	33·28 ± ·191	2·323 ± ·135	—	6·981 ± ·409	—	—

* $\underline{100\sqrt{(\text{observed s. d.})^2-(\text{simple sampling s. d.})^2}}$.

TABLE I (*b*). *Geographical Variation in Infant Mortality by Causes.*

52 Registration Counties[†] (1911—1923 inclusive).

ENGLAND AND WALES.

Group		Mean mortality per 1000	Standard Deviation		Coefficient of Variation		* (see below)
			Observed per 1000	Simple Sampling per 1000	Observed	Simple Sampling	
All Causes:							
Rural	Male	84·13 ±1·36	14·53 ± ·961	2·28	17·27 ± 1·18	2·71	17·06
	Female	65·58 ±1·08	11·59 ± ·767	2·08	17·68 ± 1·21	3·17	17·39
Urban	Male	95·13 ±1·46	15·59 ±1·03	2·64	16·39 ± 1·11	2·77	16·15
	Female	74·21 ±1·20	12·85 ± ·850	2·41	17·32 ± 1·18	3·25	17·01
Measles:							
Rural	Male	·931 ± ·054	·574 ± ·038	·250	61·64 ± 5·41	26·89	55·50
	Female	·778 ± ·045	·477 ± ·032	·235	61·30 ± 5·37	30·13	53·35
Urban	Male	1·615 ± ·066	·703 ± ·047	·361	43·53 ± 3·38	22·36	37·35
	Female	1·391 ± ·053	·572 ± ·038	·343	41·10 ± 3·15	24·63	32·90
Whooping Cough:							
Rural	Male	2·969 ± ·070	·747 ± ·049	·447	25·16 ± 1·77	15·04	20·16
	Female	3·371 ± ·064	·685 ± ·045	487	20·318 ± 1·40	14·46	14·29
Urban	Male	2·957 ± ·074	·794 ± ·053	·488	26·854 ± 1·90	16·51	21·18
	Female	3·528 ± ·083	·890 ± ·059	·545	25·23 ± 1·77	15·45	19·94
Diphtheria and Croup:							
Rural	Male	·117 ± ·009	·097 ± ·006	·089	83·08 ± 8·48	75·78	32·97
	Female	·095 ± ·008	·081 ± ·005	·082	85·30 ± 8·84	86·24	imaginary
Urban	Male	·193 ± ·012	·132 ± ·009	·125	68·48 ± 6·30	64·66	21·97
	Female	·140 ± ·014	·147 ± ·010	·109	105·07 ±12·45	77·68	70·45
Tuberculosis (all kinds):							
Rural	Male	1·655 ± ·054	·575 ± ·038	·334	34·77 ± 2·56	20·16	28·28
	Female	1·292 ± ·039	·421 ± ·028	·302	32·59 ± 2·37	23·39	22·70
Urban	Male	2·321 ± ·069	·735 ± ·049	·433	31·68 ± 2·30	18·64	25·59
	Female	1·822 ± ·077	·821 ± ·054	·392	45·05 ± 3·53	21·52	39·59
Respiratory Diseases (other than Tuberculosis):							
Rural	Male	14·80 ± ·338	3·61 ± ·239	·991	24·41 ±1·71	6·70	23·45
	Female	11·07 ± ·271	2·89 ± ·191	·880	26·13 ±1·84	7·95	24·87
Urban	Male	17·83 ± ·404	4·32 ± ·286	1·19	24·21 ±1·69	6·67	23·29
	Female	13·15 ± ·333	3·56 ± ·236	1·05	27·09 ±1·92	7·97	25·87
Diarrhoea:							
Rural	Male	7·416 ± ·268	2·861 ± ·189	·704	38·59 ±2·91	9·50	37·39
	Female	5·579 ± ·218	2·329 ± ·154	·626	41·74 ±3·21	11·23	40·21
Urban	Male	11·100 ± ·311	3·320 ± ·220	·942	29·91 ±2·15	8·49	28·69
	Female	8·472 ± ·305	3·260 ± ·216	·843	38·48 ±2·90	9·95	37·18
Congenital Debility:							
Rural	Male	35·30 ± ·364	3·893 ± ·258	1·52	11·026 ± ·738	4·29	10·14
	Female	28·07 ± ·318	3·403 ± ·225	1·39	12·124 ± ·814	4·95	11·05
Urban	Male	35·23 ± ·492	5·261 ± ·348	1·66	14·932 ±1·009	4·71	14·17
	Female	29·28 ± ·395	4·223 ± ·279	1·55	14·421 ± ·973	5·29	13·41

* $100\,\dfrac{\sqrt{(\text{observed s. d.})^2-(\text{simple sampling s. d.})^2}}{\text{mean}}$.

[†] For grouping of smaller counties see footnote on p. 331.

TABLE I (*c*). *Variation by Occupation of Father*, 1911 (*Legitimate only*).

92 Occupations with over 1000 male and 1000 female births.

All Causes	Male	129·65 ±2·10	29·88 ±1·49	7·30	23·05 ±1·21	5·63	22·35
	Female	106·50 ±1·88	26·67 ±1·33	6·82	25·04 ±1·32	6·41	24·21

and shows that out of the 33 cases which do not overlap, we get 17 cases where there still appears a real excess of female variation and 16 where there is equality or a male excess. The only cases where we never find a male excess are All Causes, and Diarrhoea, and the only case where we never find a female excess is Whooping Cough.

These cases are not all equally fitted for the application of Lenz's argument, but we shall return to this point later, and for the present consider simply the numerical values apart from the causes. Now an excess female relative variation is not enough to produce a negative correlation, unless it is large enough for $\frac{\sigma_x^2}{\bar{x}} - \frac{\sigma_y^2}{\bar{y}}$ to be also negative, or if positive, $< r_{xy}\,\sigma_x\sigma_y\left(\frac{1}{\bar{x}} - \frac{1}{\bar{y}}\right)$. Table II gives, in the last three columns but one, the observed and simple sampling values of $\frac{\sigma_x^2}{\bar{x}^2} - \frac{\sigma_y^2}{\bar{y}^2}$, and also the sign of r deduced from the approximate formula (1). Simple sampling alone would, as we have pointed out before, give in every case a negative sign for the correlation, except in that of whooping cough where it would be positive by virtue of the predominating term $r_{xy}\,\sigma_x\sigma_y\left(\frac{1}{\bar{x}} - \frac{1}{\bar{y}}\right)$. The signs of r marked with an asterisk in Table II are the cases where both factors of the numerator have the same sign; where the signs differ we have had to approximate still further by taking $r_{xy} = \cdot 97$ (a value found by experiment). We have tested this approximation in five of the latter cases by direct calculation of r, and find the signs agree in all cases but one, where the correlation is negligible. The approximations hence seem justified.

It appears that when the signs of the two terms in the numerator differ, in rather more than 2/3 (17 as against 8) of the cases the sign of the second prevails over the first. That is to say in the large majority of cases the sign of the correlation is due simply to the fact that the mean male mortality is larger (or in the case of whooping cough smaller) than the female. It is clear also from Table II, where the causes are arranged in order of size of the absolute excess of the male mortality over the female, that the positive correlations tend on the whole to appear more frequently when this excess is small or has become a defect as in whooping cough.

This tendency, and the reversal in the case of whooping cough, suggest that the correlations (if the variation is not too large) may be thrown back simply on to the size of the excess mortality, quite irrespective of *which* sex is in excess or of the cause in question.

As we pointed out before, Lenz's argument does not apply to a single cause taken separately. He says "denn wenn die Übersterblichkeit der Knaben in wesentlichen auf der Auswirkung krankhafter Erbanlagen beruht, so wird sie in Zeiten wo die Säuglingsterblichkeit *aus äusseren Ursachen* (italics ours) besonders hoch ist, verhältnismässig niedriger sein" (pp. 128—129). But from Table II we find that the negative correlation frequently does apply to a single cause. Take

TABLE II. *Comparison of Male and Female Variation in Infant Mortality.*

x = Male Mortality per person. y = Female Mortality per person.

Group		Mean mortality per 1000	Male Mean − Female Mean	100 × Male Mean ÷ Female Mean	Male S. D. − Female S. D.	Male Coefficient of Variation − Female C. V.	$100\left(\frac{\sigma^2 x}{\bar{x}} - \frac{\sigma^2 y}{\bar{y}}\right)$ Observed	$100\left(\frac{\sigma^2 x}{\bar{x}} - \frac{\sigma^2 y}{\bar{y}}\right)$ Simple Sampling	Sign of r by approximate formula from observed values	r by direct calculation
All Causes:										
London (secular), 1839—1922	Male	157·90	25·40	(119·2)	2·14	−1·48	−·0001	−·000017*	−	
	Female	132·50								
England and Wales (secular), 1839—1922	Male	153·85	27·66	121·9	2·49	−1·28	·0108	−·00012	−	
	Female	126·19								
92 Occupations, 1911	Male	129·65	23·15	121·7	3·21	−1·99	·0207	−·00266	−	
	Female	106·50								
Registration Counties, 1911—1923, Urban	Male	95·13	20·92	128·2	2·74	− ·93	·0330	−·00051	−	
	Female	74·21								
" " " Rural	Male	84·13	18·55	128·3	2·94	− ·41	·0461	−·00044	−	
	Female	65·58								
Congenital Debility:										
England and Wales, 1847—1913	Male	(45·32)	(7·84)	(120·9)	(·36)	(− ·28)	...	...	...	
	Female	(37·48)								
England and Wales, 1847—1922	Male	44·43	7·85	121·5	·22	−1·12	−·0020	−·000012*	−	
	Female	36·58								
London, 1847—1913	Male	(39·94)	(7·66)	(123·7)	(·898)	(1·08)	...	...	...	
	Female	(32·28)								
London, 1847—1922	Male	(39·29)	(6·77)	(120·8)	(·533)	(− ·29)	...	...	...	
	Female	(32·52)								
London, 1845—1922	Male	39·03	6·67	120·6	·695	·05	·0040	−·00080	−	·012 ± ·076
	Female	32·36								
Registration Counties, 1911—1923, Rural	Male	35·30	7·23	125·8	·49	−1·098	·0017	−·00038	−	
	Female	28·07								
" " " Urban	Male	35·23	5·95	120·3	1·038	·511	·0176	−·00041	+	
	Female	29·28								
Respiratory Diseases other than Tuberculosis:										
London, 1847—1913	Male	(31·37)	(6·23)	(124·8)	(1·08)	(·632)	...	...	...	
	Female	(25·14)								
London, 1847—1922	Male	(30·23)	(6·02)	(124·9)	(1·18)	(·365)	...	...	...	
	Female	(24·21)								
London, 1845—1922	Male	30·19	6·058	125·1	1·15	·246	·0218	−·00008	+	·067 ± ·076
	Female	24·13								
England and Wales, 1847—1922	Male	26·63	5·871	128·3	·89	− ·202	·0134	−·000009	−	
	Female	20·76								
Registration Counties, 1911—1923, Urban	Male	17·83	4·68	135·6	·76	−2·88	·0080	−·00040	−	
	Female	13·15								
" " " Rural	Male	14·80	3·73	133·7	·72	−1·72	·0126	−·00036	−	
	Female	11·07								

London, 1845—1922	Male Female	22·506 19·424	3·082	115·9	·489	– 2·753	– ·0030	– ·00008*	–	
England and Wales, 1847—1922	Male Female	19·080 16·098	2·982	118·5	·652	– 3·141	·0041	– ·00001	–	– ·549 ± ·094
Registration Counties, 1911—1923, Urban	Male Female	11·100 8·472	2·628	131·0	·060	– 8·57	– ·0261	– ·00038*	–	
” ” ” Rural	Male Female	7·416 5·579	1·837	132·9	·532	– 3·150	·0132	– ·00034	–	
Tuberculosis (*all*):										
London, 1845—1922	Male Female	11·68 9·44	2·24	123·7	– ·130	– 11·13	– ·0602	– ·00008*	–	
England and Wales, 1847—1922	Male Female	8·465 6·768	1·697	125·1	·591	– ·50	·0205	– ·000011	–	
Registration Counties, 1911—1923, Urban	Male Female	2·321 1·822	·499	127·4	– ·086	– 13·37	– ·0137	– ·00037*	–	
” ” ” Rural	Male Female	1·655 1·292	·363	128·1	·154	2·18	·0063	– ·00033	+	
Measles:										
London, 1845—1922	Male Female	3·48 2·99	·49	116·4	·144	– ·52	·0042	– ·00007	–	
England and Wales, 1847—1913	Male Female	(2·74) (2·35)	(·39)	(116·6)	(·127)	(1·58)	…	…	…	
England and Wales, 1847—1922	Male Female	2·66 2·28	·38	116·7	·137	·41	·0046	– ·000010	+	·256 ± ·072
Registration Counties, 1911—1923, Urban	Male Female	1·615 1·391	·224	116·1	·131	2·43	·0071	– ·00037	+	
” ” ” Rural	Male Female	·931 ·778	·153	119·7	·097	·34	·0061	– ·00034	+	
Diphtheria and Croup:										
London, 1859—1922	Male Female	1·094 ·851	·240	128·6	·171	5·17	·0111	·00000	+	
London, 1866—1922	Male Female	(·961) (·769)	(·192)	(125·0)	(·097)	(1·24)	…	…	…	
England and Wales, 1855—1922	Male Female	·865 ·668	·197	129·5	·130	·93	·0086	– ·00001	+	
England and Wales, 1866—1922	Male Female	(·686) (·528)	(·157)	(129·8)	(·091)	(2·24)	…	…	…	
Registration Counties, 1911—1923, Urban	Male Female	·193 ·140	·053	137·9	– ·015	– 36·59	– ·0064	– ·00037*	–	
” ” ” Rural	Male Female	·117 ·095	·022	123·2	·016	– 2·22	·0012	– ·00033	–	
Whooping Cough:										
London, 1845—1922	Male Female	6·64 7·31	– ·67	90·8	– ·230	·22	– ·0078	– ·00007	+	·202 ± ·074
England and Wales, 1847—1913	Male Female	(5·43) (6·10)	(– ·67)	(89·0)	(– ·078)	(·78)	…	…	…	
England and Wales, 1847—1922	Male Female	5·16 5·80	– ·64	89·0	– ·11	·74	– ·0018	– ·00001	+	
Registration Counties, 1911—1923, Rural	Male Female	2·969 3·371	– ·402	88·1	·062	4·84	·0049	– ·00033*	+	
” ” ” Urban	Male Female	2·957 3·528	– ·571	83·8	– ·096	1·62	– ·0011	– ·00036	+	

diarrhoea, for instance, the sign from formula (1) is always negative, and in the case checked $r = -\cdot549 \pm \cdot094$, a fairly substantial value. Here we have a case where the "äusseren Ursachen" are at their height, and the "Übersterblichkeit" of the boys from these same "äusseren Ursachen" is relatively low, which cannot be explained by Lenz's theory.

If we try to classify the different causes according to whether by formula (1) they would give a negative or positive correlation we find that the sign is not always the same for the different series treated under each cause. Putting them in order of the stability of the negative signs we get (omitting overlapping series):

(1) All negative.	Diarrhoea and All Causes.
(2) ¾ negative.	Tuberculosis, Congenital Debility and Respiratory Diseases.
(3) ½ negative and ½ positive.	Diphtheria and Croup.
(4) ¾ positive.	Measles.
(5) All positive.	Whooping Cough.

If we classify them according to the average excess of the female coefficient of variation over the male we get:

(1) Diphtheria and Croup.
(2) Tuberculosis.
(3) Diarrhoea.
(4) All Causes.
(5) Congenital Debility.
(6) Respiratory Diseases.
Male excess (7) Measles.
Male excess (8) Whooping Cough.

If we finally classify them according to that part of the excess of the female coefficient of variation over the male which is not due to simple sampling (last col. of Table I) we get:

(1) Diphtheria and Croup.
(2) Tuberculosis.
(3) Diarrhoea.
(4) All Causes.
(5) Respiratory Diseases.
(6) Congenital Debility.
Male excess (7) Whooping Cough.
Male excess (8) Measles.

This average classification may be to some extent misleading, as the variations are of different type and the geographical ones more unstable owing to smaller numbers (the diphtheria and croup group for instance owes its high position in the last two classifications to a single large female coefficient of variation in the Urban Counties with a large probable error*). In other cases too, different choice of groupings or of periods covered give different results as regards the relative size of the male and female coefficients of variation. We must therefore base our judgment on the most consistent results. The causes in which the negative correlation and the female excess relative variation are most stable are Diarrhoea, All Causes and then Tuberculosis, and those with a positive correlation and male excess relative variation are Whooping Cough and Measles. It is not easy from these results to find support for any theory of connection between sex differences in variation and unavoidable mortality attached more to one sex than to the other which might be supposed to occur in groups such as All Causes or Congenital Diseases, as opposed to the more definitely epidemic causes.

Consideration of the secular graphs of the course of the mortality and of the percentage $\frac{\text{male}}{\text{female}}$ mortality both for England and Wales and for London, shows that in the causes with a small absolute rate such as measles, and still more diphtheria and croup, as might be expected, the fluctuations in the $\frac{\text{male}}{\text{female}}$ mortality are so large as to mask any general tendency that might be present. In the causes with large mortality, the excess variation is clearly due mainly but not solely to the improvement in infant mortality since 1900.

The general impression given both by the tables and the graphs (which are not reproduced) is that it is a general rule for a greater absolute variation and a less relative variation to go (in any cause or group of causes) with that sex that has the higher mortality in that cause, irrespective of which sex it is, but that in causes where the absolute mortality is lowest this tendency is obscured or overcome so far as regards the relative variation. It is unfortunate that whooping cough is the only cause of death which persistently strikes female infants more than male, so that this cannot be tested further in that direction.

To a certain extent these results are unsatisfactory as they do not appear to afford a basis for any theory which would give us a clue as to what proportion of infant deaths are to any extent "inevitable" in any cause. We do not therefore feel justified in doing more than simply presenting the facts of this analysis, which do not in our opinion uphold Lenz's theory, without attempting to come to any definite conclusion as to inherent differences in male and female variation.

---

* We cannot from these data calculate with any sort of accuracy the probable error of the difference between the male and the female coefficients of variation, as without making a probably quite unjustified assumption of normality, we do not know the correlation in error between the male and female coefficients of variability.

APPENDIX. TABLE A.

Diseases included in the various groups, according to the headings in the Registrar General's Annual Report			Numbers refer to the Registrar General's Short List	
1845—1880	**1881—1900**	**1901—1910**	**1911—1920**	**1921—1922**
(*a*) All Causes	All Causes	All Causes	All Causes	All Causes
(*b*) Measles	Measles	Measles	3	3
(*c*) Whooping Cough	Whooping Cough	Whooping Cough	5	5
(*d*) Whole of IV. 1 i.e. Congenital Malformations (1858 and before called Diseases of Growth, Nutrition and Decay) and Developmental Diseases of children also IV. 4 Atrophy and Debility	Premature Birth Spina bifida Imperforate anus Cleft Palate hare-lip Other Congenital Defects Cyanosis Debility, Atrophy, Inanition i.e. Developmental Diseases V, omitting Atelectaris and adding Debility, Atrophy and Inanition	Premature Birth Congenital Hydrocephalus Other Congenital Defects Want of Breast Milk Atrophy. Debility	25	28
(*e*) Diphtheria Cynanche Maligna Croup	Diphtheria and Croup	Diphtheria and Croup (not spasmodic nor membranous)	6	6
(*f*) Whole of Group II. 2 i.e. Tubercular Diseases	Tabes Mesenterica Tubercular Meningitis (Acute Hydrocephalus) Phthisis Other forms of Tuberculosis. Scrofula Lupus	Pulmonary Tuberculosis (Tuberculous Phthisis) Phthisis (not otherwise defined) Tuberculous Meningitis " Peritonitis Tabes Mesenterica Lupus Tubercle of other organs General Tuberculosis Scrofula	9—11	10 and 11
(*g*) Whole of Group III. 3 i.e. Diseases of the Respiratory Organs	Laryngitis Other Diseases of Larynx and Trachea Emphysema, Asthma Bronchitis Pneumonia Pleurisy Other and Undefined Diseases of the Respiratory System i.e. whole of Group VI. 4, omitting Croup and adding Laryngismus Stridulus	Pneumonia: Lobar, Broncho, Epidemic, Not defined Laryngismus Stridulus Laryngitis Membranous Laryngitis (not Diphtheric) Other Diseases of Larynx and Trachea Bronchitis Emphysema Asthma Pleurisy Fibroid Disease of Lung Other Diseases of the Respiratory System	16—18	18—20
(*h*) Dysentery	Cholera Diarrhoea, Dysentery and Enteritis and Gastro-enteritis i.e. Group I. 2 and Enteritis Before 1899 enteritis and gastro-enteritis were not distinguished	Diarrhoea due to food Infective Enteritis Epidemic Diarrhoea Diarrhoea (not otherwise defined) Dysentery Enteritis (not epidemic) Gastro-enteritis	19	22

TABLE OF THE FIRST TWENTY TETRACHORIC FUNCTIONS TO SEVEN DECIMAL PLACES.

BY ALICE LEE, D.Sc.

(*With an Introductory Note by* KARL PEARSON.)

THE present table was computed as ancillary to a complete table for tetrachoric coefficients of correlation, which will shortly be published. It gives the tetrachoric functions from 0 to 19 to seven decimal places for argument intervals of $h\,(=x/\sigma_x)$ equal to ·1 and proceeds from 0 to 4·0.

The Biometric School defines the tetrachoric function $\tau_s(h)$ by the equation:

$$\tau_s(h) = \frac{(-1)^{s-1}}{\sqrt{s!}} \frac{d^{s-1}}{dh^{s-1}} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2} \right) \text{......................}(1).$$

In place of the tetrachoric function the Scandinavians have used the derivatives of:

$$\phi(h) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2}.$$

This suffers from certain disadvantages; it gives no simple nomenclature for the integral $\int_h^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$ which we take as τ_0 or $\frac{1}{2}(1-\alpha_h)$ in Sheppard's notation; and further it leads to much range of variation in the functions themselves, so that the differences may be very considerable at one part and small at another part of the table. The simple expedient of using (1) as our function causes all the tetrachorics to be proper fractions, less than unity. Further (1) is the form which arises naturally, when we are dealing with tetrachoric coefficients of correlation, and saves then a large amount of labour.

In the *Tables for Statisticians and Biometricians** a table (XXIX) is given providing the values of τ_1, τ_2, τ_3, τ_4, τ_5 and τ_6 for each one thousandth in the value of $\frac{1}{2}(1-\alpha_h)$. The value of h is also given, but not of course by equal increments of argument. A useful table of $\phi(t) \doteq \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$, and its first eight differentials for the argument t proceeding by hundredths, is given by Professor James W. Glover, pp. 392—411 of his *Tables of Applied Mathematics in Finance, Insurance and Statistics*. But this table only goes to five decimal places, and third differences would be requisite to use the table to its maximum advantage. The ideal table would be one of tetrachoric functions going to 7 or 8 decimals and exhibiting first and second differences, the argument proceeding by ·01. Our present table forms a framework for the construction of such a table. But while the computing would be possible the cost of printing the requisite number of royal octavo pages of figures would at present be prohibitive. Accordingly we have had to content ourselves with the more modest table now printed. This table, if used to its maximum value, i.e. to obtain the tetrachoric functions to seven figures, requires the use of $\delta^2\tau_s$ and $\delta^4\tau_s$, if central difference interpolation be adopted. These have not been tabled as it would have doubled the cost of printing. They can, however, be determined with a moderate amount of labour in two different manners.

---

* **Part I issued by the Biometric Laboratory, University College, London, 2nd Edition, 1924.**

(i) Using the formulae

$$\delta^2\tau_s(h) = \tau_s(h+1) + \tau_s(h-1) - 2\tau_s(h) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(i)},$$

$$\delta^4\tau_s(h) = \delta^2\tau_s(h+1) + \delta^2\tau_s(h-1) - 2\delta^2\tau_s(h) \ldots\ldots\ldots\ldots\ldots\text{(ii)},$$

we find $\delta^2\tau_s(h)$, $\delta^2\tau_s(h+1)$, $\delta^4\tau_s(h)$ and $\delta^4\tau_s(h+1)$ for interpolating a value between $\tau_s(h)$ and $\tau_s(h+1)$. This involves taking out $\tau_s(h-2)$, $\tau_s(h-1)$, $\tau_s(h)$, $\tau_s(h+1)$, $\tau_s(h+2)$ and $\tau_s(h+3)$, i.e. six entries in all.

For example, suppose we desire a value of τ_{13} between 1 and 1·1. We have:

$$\begin{array}{lll}
\tau_{13}(0{\cdot}8) = -{\cdot}042{,}6729, & & \\
\tau_{13}(0{\cdot}9) = -{\cdot}043{,}0704, & \delta^2\tau_{13}(0{\cdot}9) = +{\cdot}005{,}3224, & \\
\tau_{13}(1{\cdot}0) = -{\cdot}038{,}1455, & \delta^2\tau_{13}(1{\cdot}0) = +{\cdot}004{,}1923, & \delta^4\tau_{13}(1{\cdot}0) = -{\cdot}000{,}4818, \\
\tau_{13}(1{\cdot}1) = -{\cdot}029{,}0283, & \delta^2\tau_{13}(1{\cdot}1) = +{\cdot}002{,}5804, & \delta^4\tau_{13}(1{\cdot}1) = -{\cdot}000{,}1935, \\
\tau_{13}(1{\cdot}2) = -{\cdot}017{,}3307, & \delta^2\tau_{13}(1{\cdot}2) \quad +{\cdot}000{,}7750, & \\
\tau_{13}(1{\cdot}3) = -{\cdot}004{,}8581. & &
\end{array}$$

But the central difference formula runs, if $\phi = 1 - \theta$:

$$\tau_{13}(1{\cdot}0+\theta) = \theta\tau_{13}(1{\cdot}1) + \phi\tau_{13}(1{\cdot}0) - \frac{\theta\phi}{6}\{(1+\theta)\,\delta^2\tau_{13}(1{\cdot}1) + (1+\phi)\,\delta^2\tau_{13}(1{\cdot}0)\}$$
$$+\frac{\theta(1+\theta)\phi(1+\phi)}{120}\{(2+\theta)\,\delta^4\tau_{13}(1{\cdot}1) + (2+\phi)\,\delta^4\tau_{13}(1{\cdot}0)\} - \text{etc.} \quad \ldots\ldots(3).$$

Suppose $\theta = {\cdot}3467$, then

$$\begin{aligned}
\tau_{13}(1{\cdot}03467) = {\cdot}3467\,(-{\cdot}029{,}0283) &+ {\cdot}6533\,(-{\cdot}038{,}1455) \\
&- \tfrac{1}{6}({\cdot}3467)({\cdot}6533)\{1{\cdot}3467\,({\cdot}002{,}5804) + 1{\cdot}6533\,({\cdot}004{,}1923)\} \\
&+ \tfrac{1}{120}({\cdot}3467)(1{\cdot}3467)({\cdot}6533)(1{\cdot}6533)\{2{\cdot}3467\,(-{\cdot}000{,}1935) \\
&\qquad\qquad + 2{\cdot}6533\,(-{\cdot}000{,}4818)\} \\
&\left.\begin{array}{l} -{\cdot}010{,}0641{,}1 \\ -{\cdot}024{,}9204{,}6 \\ -{\cdot}000{,}3928{,}3 \\ -{\cdot}000{,}0072{,}8 \end{array}\right\} = -{\cdot}035{,}3847,
\end{aligned}$$

the required value.

(ii) In the second method we make use of very close expressions deduced from Taylor's Theorem for the higher differences. We have very approximately:

$$\delta^2\tau_s = \frac{1}{100}\sqrt{(s+1)(s+2)}\,\tau_{s+2} + \frac{1}{120{,}000}\sqrt{(s+1)(s+2)(s+3)(s+4)}\,\tau_{s+4},$$

or: $\delta^2\tau_s = {}_1\chi_s\, \tau_{s+2} + {}_2\chi_s\, \tau_{s+4}$ ....(4),

$$\delta^4\tau_s = \frac{1}{10{,}000}\sqrt{(s+1)(s+2)(s+3)(s+4)}\,\tau_{s+4}$$
$$+\frac{1}{6{,}000{,}000}\sqrt{(s+1)(s+2)(s+3)(s+4)(s+5)(s+6)}\,\tau_{s+6},$$

or: $\delta^4\tau_s = {}_3\chi_s\, \tau_{s+4} + {}_4\chi_s\, \tau_{s+6}$ ....(5).

Hence, if a table of the ${}_t\chi_s$'s be provided, we can write down the values of $\delta^2\tau_s$ and $\delta^4\tau_s$ by a single operation on the machine for each. This is shorter than deducing the differences from the main table itself. The formulae (4) and (5) give the differences extremely closely, so that we can proceed by this method as far as $\delta^2\tau_{13}$ and $\delta^4\tau_{13}$, not further, because we should require higher tetrachorics than τ_{19}, which is the last tabled. But most computers will rarely need to go even as far as τ_{13}. It is only in special table-construction work, for which the present table was ancillary, that these high tetrachoric functions are needful, a point to which we will recur shortly.

TABLE I.

Table of the χ-Coefficients.

Order of Tetrachoric Function s	$\delta^2\tau_s$		$\delta^4\tau_s$	
	${}_1\chi_s$	${}_2\chi_s$	${}_3\chi_s$	${}_4\chi_s$
0	·0141,4214	·0000,4082	·0004,8990	·0000,0447
1	·0244,9490	·0000,9129	·0010,9545^{-}	·0000,1183
2	·0346,4102	·0001,5811	·0018,9737	·0000,2366
3	·0447,2136	·0002,4152	·0028,9828	·0000,4099
4	·0547,7226	·0003,4157	·0040,9878	·0000,6481
5	·0648,0741	·0004,5826	·0054,9909	·0000,9612
6	·0748,3315^{-}	·0005,9161	·0070,9930	·0001,3594
7	·0848,5281	·0007,4162	·0088,9944	·0001,8526
8	·0948,6833	·0009,0830	·0108,9954	·0002,4507
9	·1048,8089	·0010,9163	·0130,9962	·0003,1639
10	·1148,9125^{+}	·0012,9164	·0154,9968	·0004,0020
11	·1248,9996	·0015,0831	·0180,9972	·0004,9751
12	·1349,0738	·0017,4165^{-}	·0208,9976	·0006,0933
13	·1449,1377	·0019,9165^{-}	·0238,9979	·0007,3664
14	·1549,1933	·0022,5832	·0270,9981	·0008,8045^{+}
15	·1649,2422	·0025,4165^{+}	·0304,9984	·0010,4177

$$\tau_s(h+\theta) = \theta\tau_s(h+1) + \phi\tau_s(h) - \tfrac{1}{6}\theta\phi\,\{(1+\theta)\,\delta^2\tau_s(h+1) + (1+\phi)\,\delta^2\tau_s(h)\}$$
$$+\tfrac{1}{120}\theta(1+\theta)\,\phi(1+\phi)\,\{(2+\theta)\,\delta^4\tau_s(h+1) + (2+\phi)\,\delta^4\tau_s(h)\} - \dots \quad \text{......(6)}.$$

$\delta^2\tau_s(h) = {}_1\chi_s \cdot \tau_{s+2}(h) + {}_2\chi_s \cdot \tau_{s+4}(h)$ ....(7).

$\delta^4\tau_s(h) = {}_3\chi_s \cdot \tau_{s+4}(h) + {}_4\chi_s \cdot \tau_{s+6}(h)$ ....(8).

For the purposes of calculation it is convenient to note that

$$\tfrac{1}{120}\theta(1+\theta)\,\phi(1+\phi) = \tfrac{1}{10}\left(\tfrac{1}{6}\theta\phi\right)\left(1+\tfrac{1}{2}\theta\phi\right),$$

which needs only one multiplication.

We will now illustrate the method of Equations (7) and (8) on the previous example.

$$\delta^2\tau_{13}(1{\cdot}0) = {\cdot}1449{,}1377\,({\cdot}029{,}2108) + {\cdot}0019{,}9165\,(-\,{\cdot}020{,}5443),$$

$$\delta^2\tau_{13}(1{\cdot}1) = {\cdot}1449{,}1377\,({\cdot}017{,}9171) + {\cdot}0019{,}9165\,(-\,{\cdot}008{,}0852).$$

Hence: $\delta^2\tau_{13}(1{\cdot}0) = {\cdot}004{,}1921$, $\delta^2\tau_{13}(1{\cdot}1) = {\cdot}002{,}5803$.

These values are as correct as, possibly more correct than, the values ·004,1923 and ·002,5804 found from the main table itself, as in the latter case the apparent higher differences may be considerably affected by the raising of the last figure of the tabled entries. Similarly

$$\delta^4\tau_{13}(1{\cdot}0) = {\cdot}0238{,}9979\,(-\,{\cdot}020{,}5443) + {\cdot}0007{,}3664\,(+\,{\cdot}012{,}5634),$$

$$\delta^4\tau_{13}(1{\cdot}1) = {\cdot}0238{,}9979\,(-\,{\cdot}008{,}0852) + {\cdot}0007{,}3664\,(-\,{\cdot}000{,}2002).$$

Hence: $\delta^4\tau_{13}(1{\cdot}0) = -\,{\cdot}000{,}4817$, $\delta^4\tau_{13}(1{\cdot}1) = -\,{\cdot}000{,}1934$.

These are again in good agreement with the values found, i.e. − ·000,4818 and − ·000,1935, by the more laborious operation from the main table.

If we again substitute in equation (6) to find $\tau_{13}(1{\cdot}03467)$ we have

$$\tau_{13}(1{\cdot}03467) = -\,{\cdot}035{,}3847$$

as before*.

We may conclude, I think, that the second method is shorter and fully adequate if we do not want differences beyond those of τ_{13}. If we do, we must either compute enough additional τ_s's by means of the difference formula:

$$\tau_s(h) = hp_s\tau_{s-1}(h) - q_s\tau_{s-2}(h) \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9),$$

where p_s and q_s are given in Table II below, or we must of necessity fall back on the first method. The first method, however, will fail whatever be the value of s

TABLE II.

Values of p_s and q_s.

s	p_s	q_s	s	p_s	q_s
2	·707,1067,812	·000,0000,000	14	·267,2612,419	·889,4991,800
3	·577,3502,692	·408,2482,905	15	·258,1988,897	·897,0852,271
4	·500,0000,000	·577,3502,692	16	·250,0000,000	·903,6961,141
5	·447,2135,955$^-$	·670,8203,933	17	·242,5356,250	·909,5085,939
6	·408,2482,905$^-$	·730,2967,433	18	·235,7022,604	·914,6591,208
7	·377,9644,730	·771,5167,498	19	·229,4157,339	·919,2547,198
8	·353,5533,906	·801,7837,257	20	·223,6067,978	·923,3805,169
9	·333,3333,333	·824,9579,114	21	·218,2178,902	·927,1050,693
10	·316,2277,660	·843,2740,427	22	·213,2007,163	·930,4842,104
11	·301,5113,446	·858,1163,303	23	·208,5144,141	·933,5638,714
12	·288,6751,346	·870,3882,798	24	·204,1241,452	·936,3821,838
13	·277,3500,981	·880,7048,459	25	·200,0000,000	·938,9710,681

N.B. The last figure in q_s may be a unit in doubt, but this will not affect seven figure accuracy in the τ's.

---

[*] Actually by the first method we have − ·035,3846,8 and by the second method − ·035,3846,6.

when we approach the under limit of our main table, for we cannot get $\delta^4\tau_s(h)$ from the table itself for values of h greater than 3·7, for $\delta^4\tau_s(h)$ requires a knowledge of $\tau_s(h+3)$; we are therefore compelled in such cases to work with the less accurate backward difference method. Our second method does not however fail in these cases, even for $h=4$, until $s=13$, and accordingly should be used at the foot of the table for $h=3·8$, $3·9$ and $4·0$ for values of s from 0 to 13. We have accordingly the following scheme:

Order of Tetrachoric Functions.

<table>
<tr><th>h s=</th><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th><th>16</th><th>17</th><th>18</th><th>19</th></tr>
<tr><td>0·0</td><td rowspan="38" colspan="14">Central Differences,
Second Method</td><td rowspan="38" colspan="6">Central Differences,
First Method</td></tr>
<tr><td>0·1</td></tr>
<tr><td>0·2</td></tr>
<tr><td>0·3</td></tr>
<tr><td>0·4</td></tr>
<tr><td>0·5</td></tr>
<tr><td>0·6</td></tr>
<tr><td>0·7</td></tr>
<tr><td>0·8</td></tr>
<tr><td>0·9</td></tr>
<tr><td>1·0</td></tr>
<tr><td>1·1</td></tr>
<tr><td>1·2</td></tr>
<tr><td>1·3</td></tr>
<tr><td>1·4</td></tr>
<tr><td>1·5</td></tr>
<tr><td>1·6</td></tr>
<tr><td>1·7</td></tr>
<tr><td>1·8</td></tr>
<tr><td>1·9</td></tr>
<tr><td>2·0</td></tr>
<tr><td>2·1</td></tr>
<tr><td>2·2</td></tr>
<tr><td>2·3</td></tr>
<tr><td>2·4</td></tr>
<tr><td>2·5</td></tr>
<tr><td>2·6</td></tr>
<tr><td>2·7</td></tr>
<tr><td>2·8</td></tr>
<tr><td>2·9</td></tr>
<tr><td>3·0</td></tr>
<tr><td>3·1</td></tr>
<tr><td>3·2</td></tr>
<tr><td>3·3</td></tr>
<tr><td>3·4</td></tr>
<tr><td>3·5</td></tr>
<tr><td>3·6</td></tr>
<tr><td>3·7</td></tr>
<tr><td>3·8</td><td rowspan="3" colspan="14">Central Differences,
Second Method</td><td rowspan="3" colspan="6">Backward Differences,
First Method</td></tr>
<tr><td>3·9</td></tr>
<tr><td>4·0</td></tr>
</table>

By the "First Method" is merely meant finding any differences from the table itself, and by the "Second Method" finding central differences from the formulae

(7) and (8). It is open to question, whether for τ_{14} and τ_{15} the use of (9) might not be better, at any rate for the range $h = 3{\cdot}8$ to $4{\cdot}0$ inclusive.

It may be desirable to give further illustration. Let us suppose we need to find $\tau_8(1{\cdot}4765)$.

Here $\theta = {\cdot}765$, $\phi = {\cdot}235$.

$$\begin{aligned}
\delta^2\tau_8(1{\cdot}4) &= {\cdot}0948{,}6833\,(-{\cdot}036{,}1788) + {\cdot}0009{,}0830\,({\cdot}034{,}5869),\\
\delta^2\tau_8(1{\cdot}5) &= {\cdot}0948{,}6833\,(-{\cdot}036{,}7973) + {\cdot}0009{,}0830\,({\cdot}030{,}4286),\\
\delta^4\tau_8(1{\cdot}4) &= {\cdot}0108{,}9954\,({\cdot}034{,}5869) + {\cdot}0002{,}4507\,(-{\cdot}028{,}2655),\\
\delta^4\tau_8(1{\cdot}5) &= {\cdot}0108{,}9954\,({\cdot}030{,}4286) + {\cdot}0002{,}4507\,(-{\cdot}020{,}6873).
\end{aligned}$$

Hence:

$$\begin{aligned}
&\delta^2\tau_8(1{\cdot}4) = -{\cdot}003{,}4008, \quad \delta^2\tau_8(1{\cdot}5) = -{\cdot}003{,}4633,\\
&\delta^4\tau_8(1{\cdot}4) = +{\cdot}000{,}3701, \quad \delta^4\tau_8(1{\cdot}5) = +{\cdot}000{,}3266^{*}.
\end{aligned}$$

Accordingly:

$$\begin{aligned}
\tau_8(1{\cdot}4765) &= {\cdot}765 \times ({\cdot}035{,}1482) + {\cdot}235\,({\cdot}028{,}8707)\\
&\quad - \frac{{\cdot}765 \times {\cdot}235}{6}\{1{\cdot}765\,(-{\cdot}003{,}4633) + 1{\cdot}235\,(-{\cdot}003{,}4008)\}\\
&\quad + \frac{1}{10}\,\frac{{\cdot}765 \times {\cdot}235}{6}\,(1 + \tfrac{1}{2}({\cdot}765 \times {\cdot}235))\,\{2{\cdot}765\,(+{\cdot}000{,}3266)\\
&\qquad\qquad + 2{\cdot}235\,(+{\cdot}000{,}3701)\}\\
&= \left.\begin{array}{l} {\cdot}033{,}6729{,}9\\ +{\cdot}000{,}3089{,}9\\ +{\cdot}000{,}0056{,}5 \end{array}\right\} = {\cdot}033{,}9876.
\end{aligned}$$

We will compare this with the result obtainable from Professor Glover's Table for $\phi^{(7)}(1{\cdot}4765)$†.

Clearly,
$$\begin{aligned}
\tau_8(1{\cdot}4765) &= \frac{(-1)^7}{\sqrt{8!}}\,\phi^{(7)}(1{\cdot}4765)\\
&= -\frac{\phi^{(7)}(1{\cdot}4765)}{200{\cdot}79840637}.
\end{aligned}$$

Taking out of Glover's Table:

$$\left.\begin{array}{l}
\phi^{(7)}(1{\cdot}45) = -6{\cdot}51527\\
\phi^{(7)}(1{\cdot}46) = -6{\cdot}63784\\
\phi^{(7)}(1{\cdot}47) = -6{\cdot}75337\\
\phi^{(7)}(1{\cdot}48) = -6{\cdot}86185\\
\phi^{(7)}(1{\cdot}49) = -6{\cdot}96329\\
\phi^{(7)}(1{\cdot}50) = -7{\cdot}05770
\end{array}\right\}\ \text{we find:}\quad
\begin{array}{l}
\delta^2\phi^{(7)}(1{\cdot}46) = +{\cdot}00704,\\
\delta^2\phi^{(7)}(1{\cdot}47) = +{\cdot}00705,\\
\delta^2\phi^{(7)}(1{\cdot}48) = +{\cdot}00704,\\
\delta^2\phi^{(7)}(1{\cdot}49) = +{\cdot}00703.
\end{array}$$

It is accordingly idle to proceed to $\delta^4\phi^{(7)}(h)$.

---

* The corresponding values found from the table itself are: $-{\cdot}003{,}4009$, $-{\cdot}003{,}4634$, $+{\cdot}000{,}3701$ and $+{\cdot}000{,}3268$, and the agreement is seen to be good.

† *Loc. cit.* pp. 397 and 399.

Substituting in the central difference formula (6), we have $\theta = {\cdot}65$, $\phi = {\cdot}35$ and:

$$\begin{aligned}\phi^{(7)}(1{\cdot}4765) &= {\cdot}65\,(-\,6{\cdot}86185) + {\cdot}35\,(-\,6{\cdot}75337) \\ &\quad - \tfrac{1}{6}\,({\cdot}65 \times {\cdot}35)\,\{1{\cdot}65\,({\cdot}00704) + 1{\cdot}35\,({\cdot}00705)\} \\ &= -\,6{\cdot}824{,}683.\end{aligned}$$

Accordingly: $\tau_8(1{\cdot}4765) = +\,{\cdot}033{,}9877$,

while the value found from our table is ·033,9876, a difference of only a unit in the seventh decimal. Thus we can depend on Glover's Table giving us the tetrachoric function correct to the sixth place[*]. On the other hand we find for $\phi^{(7)}(1{\cdot}4765)$ as calculated from our seven-figure tetrachoric function the value − 6·824,656, which suggests that in an interpolated value Glover's Table may give us an error of 3 in the fifth decimal place. For a great variety of statistical purposes this is ample, but it indicates that we are not introducing unnecessary figures when we supplement for many purposes our original five-figure tables of the tetrachoric functions by seven-figure tables.

Those who choose to describe frequency not diverging too greatly from the normal by tetrachoric series, will find two fundamental formulae useful. The first gives the ordinate z_x of the frequency curve in terms of the total frequency N, the standard deviation σ_x for the character x and tetrachoric functions of $h = x/\sigma_x$. The second gives the area of the frequency curve from a given value of $h = x/\sigma_x$ up to the end of the range, i.e. $h = \infty$.

They are, if we represent this tail area by $N\tfrac{1}{2}(1 - \alpha_h)$, in accordance with the notation of the probability integral of the normal curve:

$$z_x = \frac{N}{\sigma_x}\left\{\tau_1(h) + \frac{4}{\sqrt{24}}\sqrt{\beta_1}\,\tau_4(h) + \frac{5}{\sqrt{120}}(\beta_2 - 3)\,\tau_5(h)\right\} \quad \ldots\ldots\ldots(10),$$

or:

$$z_x = \frac{N}{\sigma_x}\{\tau_1(h) + {\cdot}8164{,}9658\sqrt{\beta_1}\,\tau_4(h) + {\cdot}4564{,}3546\,(\beta_2 - 3)\,\tau_5(h)\} \quad \ldots\ldots\ldots(10)\ \textit{bis}.$$

Integrating:

$$\int_{-\infty}^{h} z_x\,dx = N\left\{\tau_0'(h) - \frac{1}{\sqrt{6}}\sqrt{\beta_1}\,\tau_3(h) - \frac{1}{\sqrt{24}}(\beta_2 - 3)\,\tau_4(h)\right\},$$

or:

$$N\tfrac{1}{2}(1 + \alpha_h) = N\{\tau_0'(h) - {\cdot}4082{,}4829\sqrt{\beta_1}\,\tau_3(h) - {\cdot}2041{,}2415\,(\beta_2 - 3)\,\tau_4(h)\} \quad \ldots\ldots\ldots(11).$$

On the other hand, if h be negative, α_h changes sign and the even tetrachoric functions change sign. Thus:

$$\begin{aligned}\int_{-\infty}^{-h} z_x\,dx = N\tfrac{1}{2}(1 - \alpha_h) = N\{\tau_0''(h) &- {\cdot}4082{,}4829\sqrt{\beta_1}\,\tau_3(h) \\ &+ {\cdot}2041{,}2415\,(\beta_2 - 3)\,\tau_4(h)\} \ldots(11)\ \textit{bis}.\end{aligned}$$

---

[*] The divergence between the two tables is far from being always as great as this. Thus I found from the present table that $\tau_9(1{\cdot}4765) = -\,{\cdot}017{,}8327$, while Glover's Table gave the value $-\,{\cdot}017{,}8326$, a close accordance, but in this case the contributions of $\delta^2\tau_9(1{\cdot}4)$ and $\delta^2\tau_9(1{\cdot}5)$ only give a total when combined with those of $\delta^4\tau_9(1{\cdot}4)$ and $\delta^4\tau_9(1{\cdot}5)$ of $-\,{\cdot}000{,}0080$, so that little depends on the higher differences.

Here $\beta_1 = \mu_3^2/\mu_2^3$, $\beta_2 = \mu_4/\mu_2^2$, where μ_s is the sth moment coefficient of the distribution, and $\sqrt{\beta_1}$ follows the sign of μ_3. $\tau_0'(h)$ is the $\frac{1}{2}(1+\alpha)$, and $\tau_0''(h)$ is the $\frac{1}{2}(1-\alpha)$ of the probability integral of the normal curve. The $\tau_0(h)$ of our Table

$$= \int_{-\infty}^{-h} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \int_{z}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

$$= \tfrac{1}{2}(1-\alpha) \text{ if } h \text{ be negative}, \quad = 1 - \tfrac{1}{2}(1+\alpha) \text{ if } h \text{ be positive},$$

as in the second integral. Accordingly:

$$\tau_0'(h) = 1 - \tau_0(h), \quad \tau_0''(h) = \tau_0(h).$$

This artifice is adopted to avoid tabling $\tau_0(h)$ throughout the range from $-\infty$ to $+\infty$. The like difficulty does not occur with other even order tetrachorics, they are all odd in h, and merely change their sign with h. This point must be borne in mind when plotting z_x from Equations (10) and (10) *bis*; the second term, i.e. the one in $\tau_4(h)$, will be negative for h negative.

It is clear that Equations (11) and (11) *bis* will provide the frequency of any cell of a distribution, by subtracting $N\frac{1}{2}(1+\alpha_h)$ from $N\frac{1}{2}(1+\alpha_{h'})$, $h' > h$, if h and h' correspond to the limits of the subrange of the cell. For applications of these equations, the reader is referred to pp. 289—290, 303—304 of the current number.

I have to acknowledge a grant from the Government Grant Committee of the Royal Society, which has enabled Dr Alice Lee to devote her time to the calculation of this table and certain other tables shortly to be published.

Table of the First Twenty Tetrachoric Functions.

h	τ_0	τ_1	τ_2	τ_3	τ_4	h
0·0	+·500 0000	+·398 9423	+·000 0000	−·162 8675^{+}	+·000 0000	*0·0*
0·1	+·460 1722	+·396 9525^{+}	+·028 0688	−·160 4346	−·024 2273	*0·1*
0·2	+·420 7403	+·391 0427	+·055 3018	−·153 2568	−·047 2542	*0·2*
0·3	+·382 0886	+·381 3878	+·080 9046	−·141 6878	−·067 9635^{-}	*0·3*
0·4	+·344 5783	+·368 2701	+·104 1625^{+}	−·126 2904	−·085 3963	*0·4*
0·5	+·308 5375^{+}	+·352 0653	+·124 4739	−·107 7976	−·098 8144	*0·5*
0·6	+·274 2531	+·333 2246	+·141 3752	−·087 0646	−·107 7424	*0·6*
0·7	+·241 9637	+·312 2539	+·154 5578	−·065 0133	−·111 9887	*0·7*
0·8	+·211 8554	+·289 6916	+·163 8743	−·042 5758	−·111 6432	*0·8*
0·9	+·184 0601	+·266 0853	+·169 3356	−·020 6395^{-}	−·107 0537	*0·9*
1·0	+·158 6553	+·241 9707	+·171 0991	−·000 0000	−·098 7841	*1·0*
1·1	+·135 6661	+·217 8522	+·169 4492	+·018 6769	−·087 5592	*1·1*
1·2	+·115 0697	+·194 1861	+·164 7723	+·034 8815^{-}	−·074 2025^{-}	*1·2*
1·3	+·096 8005^{-}	+·171 3686	+·157 5287	+·048 2730	−·059 5717	*1·3*
1·4	+·080 7567	+·149 7275^{-}	+·148 2226	+·058 6809	−·044 4997	*1·4*
1·5	+·066 8072	+·129 5176	+·137 3742	+·066 0942	−·029 7424	*1·5*
1·6	+·054 7993	+·110 9208	+·125 4926	+·070 6419	−·015 9397	*1·6*
1·7	+·044 5655^{-}	+·094 0491	+·113 0547	+·072 5673	−·003 5900	*1·7*
1·8	+·035 9303	+·078 9502	+·100 4871	+·072 1980	+·006 9620	*1·8*
1·9	+·028 7166	+·065 6158	+·088 1550^{+}	+·069 9155^{-}	+·015 5234	*1·9*
2·0	+·022 7501	+·053 9910	+·076 3548	+·066 1252	+·022 0417	*2·0*
2·1	+·017 8644	+·043 9836	+·065 3123	+·061 2307	+·026 5842	*2·1*
2·2	+·013 9034	+·035 4746	+·055 1855^{+}	+·055 6126	+·029 3125^{-}	*2·2*
2·3	+·010 7241	+·028 3270	+·046 0696	+·049 6116	+·030 4550^{+}	*2·3*
2·4	+·008 1975^{+}	+·022 3945^{+}	+·038 0048	+·043 5184	+·030 2801	*2·4*
2·5	+·006 2097	+·017 5283	+·030 9859	+·037 5685^{-}	+·029 0708	*2·5*
2·6	+·004 6612	+·013 5830	+·024 9720	+·031 9405^{-}	+·027 1051	*2·6*
2·7	+·003 4670	+·010 4209	+·019 8955^{+}	+·026 7597	+·024 6389	*2·7*
2·8	+·002 5551	+·007 9155^{-}	+·015 6718	+·022 1033	+·021 8964	*2·8*
2·9	+·001 8658	+·005 9525^{+}	+·012 2063	+·018 0071	+·019 0630	*2·9*
3·0	+·001 3499	+·004 4318	+·009 4014	+·014 4744	+·016 2837	*3·0*
3·1	+·000 9676	+·003 2668	+·007 1610	+·011 4829	+·013 6641	*3·1*
3·2	+·000 6871	+·002 3841	+·005 3946	+·008 9933	+·011 2747	*3·2*
3·3	+·000 4834	+·001 7226	+·004 0195^{+}	+·006 9550^{+}	+·009 1551	*3·3*
3·4	+·000 3369	+·001 2322	+·002 9625^{-}	+·005 3122	+·007 3204	*3·4*
3·5	+·000 2326	+·000 8727	+·002 1598	+·004 0081	+·005 7671	*3·5*
3·6	+·000 1591	+·000 6119	+·001 5576	+·002 9877	+·004 4786	*3·6*
3·7	+·000 1078	+·000 4248	+·001 1114	+·002 2006	+·003 4296	*3·7*
3·8	+·000 0723	+·000 2919	+·000 7845^{-}	+·001 6019	+·002 5906	*3·8*
3·9	+·000 0481	+·000 1987	+·000 5478	+·001 1524	+·001 9310	*3·9*
4·0	+·000 0317	+·000 1338	+·000 3785^{+}	+·000 8195^{+}	+·001 4205^{+}	*4·0*

Table of the First Twenty Tetrachoric Functions—(continued).

h	τ_5	τ_6	τ_7	τ_8	τ_9	h
0·0	+·109 2548	+·000 0000	−·084 2919	−·000 0000	+·069 5373	*0·0*
0·1	+·106 5394	+·022 0425^{+}	−·081 3638	−·020 5500^{-}	+·066 4367	*0·1*
0·2	+·098 5813	+·042 5587	−·072 8400	−·039 2734	+·057 4717	*0·2*
0·3	+·085 9288	+·060 1576	−·059 4743	−·054 5416	+·043 6096	*0·3*
0·4	+·069 4420	+·073 7045^{-}	−·042 4326	−·065 0959	+·026 3256	*0·4*
0·5	+·050 2172	+·082 4144	−·023 1686	−·070 1742	+·007 4174	*0·5*
0·6	+·029 4944	+·085 9085^{+}	−·003 2732	−·069 5744	−·011 2147	*0·6*
0·7	+·008 5543	+·084 2295^{+}	+·015 6853	−·063 6520	−·027 7918	*0·7*
0·8	−·011 3820	+·077 8153	+·032 3105^{+}	−·053 2523	−·040 8554	*0·8*
0·9	−·029 2429	+·067 4365^{-}	+·045 5011	−·039 5911	−·049 4138	*0·9*
1·0	−·044 1776	+·054 1063	+·054 5340	−·024 1009	−·053 0219	*1·0*
1·1	−·055 6023	+·038 9747	+·059 1023	−·008 2639	−·051 7870	*1·1*
1·2	−·063 2204	+·023 2182	+·059 3064	+·006 5456	−·046 3071	*1·2*
1·3	−·067 0162	+·007 9380	+·055 6045^{+}	+·019 1923	−·037 5547	*1·3*
1·4	−·067 2256	−·005 9246	+·048 7307	+·028 8707	−·026 7277	*1·4*
1·5	−·064 2891	−·017 6481	+·039 5946	+·035 1482	−·015 0898	*1·5*
1·6	−·058 7935^{+}	−·026 7631	+·029 1754	+·037 9623	−·003 8219	*1·6*
1·7	−·051 4089	−·033 0572	+·018 4223	+·037 5773	+·006 0962	*1·7*
1·8	−·042 8277	−·036 5561	+·008 1718	+·034 5106	+·013 9649	*1·8*
1·9	−·033 7104	−·037 4849	−·000 9110	+·029 4428	+·019 3986	*1·9*
2·0	−·024 6434	−·036 2182	−·008 3656	+·023 1238	+·022 3172	*2·0*
2·1	−·016 1083	−·033 2243	−·013 9432	+·016 2865^{-}	+·022 9031	*2·1*
2·2	−·008 4664	−·029 0109	−·017 5912	+·009 5777	+·021 5356	*2·2*
2·3	−·001 9547	−·024 0766	−·019 4221	+·003 5107	+·018 7140	*2·3*
2·4	+·003 3069	−·018 8733	−·019 6716	−·001 5596	+·014 9806	*2·4*
2·5	+·007 3005^{-}	−·013 7793	−·018 6527	−·005 4388	+·010 8554	*2·5*
2·6	+·010 0902	−·009 0845^{+}	−·016 7122	−·008 0787	+·006 7853	*2·6*
2·7	+·011 8000	−·004 9870	−·014 1931	−·009 5502	+·003 1136	*2·7*
2·8	+·012 5914	−·001 5978	−·011 4054	−·010 0097	+·000 0666	*2·8*
2·9	+·012 6436	+·001 0474	−·008 6067	−·009 6643	−·002 2420	*2·9*
3·0	+·012 1371	+·002 9730	−·005 9930	−·008 7402	−·003 7962	*3·0*
3·1	+·011 2405^{-}	+·004 2467	−·003 6964	−·007 4562	−·004 6554	*3·1*
3·2	+·010 1022	+·004 9635^{+}	−·001 7907	−·006 0056	−·004 9287	*3·2*
3·3	+·008 8455^{+}	+·005 2310	−·000 3000	−·004 5441	−·004 7510	*3·3*
3·4	+·007 5673	+·005 1577	+·000 7897	−·003 1860	−·004 2623	*3·4*
3·5	+·006 3383	+·004 8449	+·001 5191	−·002 0048	−·003 5921	*3·5*
3·6	+·005 2061	+·004 3807	+·001 9441	−·001 0379	−·002 8493	*3·6*
3·7	+·004 1986	+·003 8375^{-}	+·002 1273	−·000 2940	−·002 1175^{+}	*3·7*
3·8	+·003 3280	+·003 2709	+·002 1303	+·000 2395^{+}	−·001 4540	*3·8*
3·9	+·002 5948	+·002 7212	+·002 0092	+·000 5887	−·000 8923	*3·9*
4·0	+·001 9914	+·002 2145^{-}	+·001 8116	+·000 7865^{-}	−·000 4459	*4·0*

Table of the First Twenty Tetrachoric Functions—(continued).

h	τ_{10}	τ_{11}	τ_{12}	τ_{13}	τ_{14}	h
0·0	+ ·000 0000	− ·059 6711	+ ·000 0000	+ ·052 5526	+ ·000 0000	0·0
0·1	+ ·019 4302	− ·056 4246	− ·018 5406	+ ·049 1792	+ ·017 8063	0·1
0·2	+ ·036 7531	− ·047 1011	− ·034 7088	+ ·039 5568	+ ·032 9879	0·2
0·3	+ ·050 1307	− ·032 8877	− ·046 4813	+ ·025 0968	+ ·043 3573	0·3
0·4	+ ·058 2237	− ·015 5684	− ·052 4749	+ ·007 8896	+ ·047 5198	0·4
0·5	+ ·060 3489	+ ·002 7329	− ·052 1325^{-}	− ·009 6364	+ ·045 0841	0·5
0·6	+ ·056 5425^{-}	+ ·019 8524	− ·045 7754	− ·025 1016	+ ·036 6919	0·6
0·7	+ ·047 5241	+ ·033 8789	− ·034 5184	− ·036 5389	+ ·023 8683	0·7
0·8	+ ·034 5706	+ ·043 3974	− ·020 0676	− ·042 6729	+ ·008 7263	0·8
0·9	+ ·019 3227	+ ·047 6462	− ·004 4394	− ·043 0704	− ·006 4111	0·9
1·0	+ ·003 5566	+ ·046 5713	+ ·010 3483	− ·038 1455^{-}	− ·019 3996	1·0
1·1	− ·011 0454	+ ·040 7759	+ ·022 5619	− ·029 0283	− ·028 6027	1·1
1·2	− ·023 0920	+ ·031 3819	+ ·030 9700	− ·017 3307	− ·033 1060	1·2
1·3	− ·031 6230	+ ·019 8312	+ ·034 9665^{-}	− ·004 8581	− ·032 7906	1·3
1·4	− ·036 1788	+ ·007 6639	+ ·034 5869	+ ·006 6802	− ·028 2655^{+}	1·4
1·5	− ·036 7973	− ·003 6934	+ ·030 4286	+ ·015 9119	− ·020 6873	1·5
1·6	− ·033 9464	− ·013 0967	+ ·023 4974	+ ·021 9615^{+}	− ·011 5098	1·6
1·7	− ·028 4107	− ·019 7937	+ ·015 0146	+ ·024 5117	− ·002 2188	1·7
1·8	− ·021 1529	− ·023 4636	+ ·006 2192	+ ·023 7693	+ ·005 9028	1·8
1·9	− ·013 1730	− ·024 1928	− ·001 8036	+ ·020 3562	+ ·011 9412	1·9
2·0	− ·005 3851	− ·022 3981	− ·008 2444	+ ·015 1529	+ ·015 4330	2·0
2·1	+ ·001 4755^{-}	− ·018 7193	− ·012 6322	+ ·009 1287	+ ·016 3598	2·1
2·2	+ ·006 9057	− ·013 8993	− ·014 8379	+ ·003 1875^{+}	+ ·015 0725^{+}	2·2
2·3	+ ·010 6507	− ·008 6728	− ·015 0285^{+}	− ·001 9486	+ ·012 1700	2·3
2·4	+ ·012 6846	− ·003 6761	− ·013 5875^{-}	− ·005 8068	+ ·008 3614	2·4
2·5	+ ·013 1683	+ ·000 6108	− ·011 0207	− ·008 1794	+ ·004 3378	2·5
2·6	+ ·012 3914	+ ·003 8914	− ·007 8646	− ·009 0984	+ ·000 6733	2·6
2·7	+ ·010 7118	+ ·006 0485^{-}	− ·004 6091	− ·008 7784	− ·002 2348	2·7
2·8	+ ·008 4999	+ ·007 1187	− ·001 6442	− ·007 5463	− ·004 1846	2·8
2·9	+ ·006 0936	+ ·007 2520	+ ·000 7673	− ·005 7698	− ·005 1544	2·9
3·0	+ ·003 7689	+ ·006 6667	+ ·002 4931	− ·003 7970	− ·005 2620	3·0
3·1	+ ·001 7239	+ ·005 6062	+ ·003 5165^{-}	− ·001 9140	− ·004 7137	3·1
3·2	+ ·000 0768	+ ·004 3036	+ ·003 9086	− ·000 3212	− ·003 7514	3·2
3·3	− ·001 1260	+ ·002 9566	+ ·003 7966	+ ·000 8710	− ·002 6089	3·3
3·4	− ·001 8961	+ ·001 7139	+ ·003 3324	+ ·001 6331	− ·001 4803	3·4
3·5	− ·002 2851	+ ·000 6710	+ ·002 6669	+ ·001 9979	− ·000 5033	3·5
3·6	− ·002 3685^{-}	− ·000 1258	+ ·001 9308	+ ·002 0386	+ ·000 2440	3·6
3·7	− ·002 2297	− ·000 6703	+ ·001 2247	+ ·001 8472	+ ·000 7372	3·7
3·8	− ·001 9493	− ·000 9856	+ ·000 6154	+ ·001 5166	+ ·000 9929	3·8
3·9	− ·001 5968	− ·001 1120	+ ·000 1379	+ ·001 1285^{+}	+ ·001 0536	3·9
4·0	− ·001 2272	− ·001 0974	− ·000 1991	+ ·000 7457	+ ·000 9742	4·0

Table of the First Twenty Tetrachoric Functions—(*continued*).

h	τ_{15}	τ_{16}	τ_{17}	τ_{18}	τ_{19}	h
0·0	−·047 1442	+·000 0000	+·042 8780	+·000 0000	−·039 4158	*0·0*
0·1	−·043 6581	−·017 1829	+·039 2907	+·016 6426	−·035 7364	*0·1*
0·2	−·033 7824	−·031 5002	+·029 1974	+·030 1883	−·025 4547	*0·2*
0·3	−·019 1556	−·040 6185^{-}	+·014 4667	+·038 1750^{+}	−·010 6712	*0·3*
0·4	−·002 1698	−·043 1604	−·002 2137	+·039 2681	+·005 6385^{-}	*0·4*
0·5	+·014 4650^{-}	−·038 9342	−·017 8775^{-}	+·033 5046	+·020 2772	*0·5*
0·6	+·028 2026	−·028 9280	−·029 8601	+·022 2364	+·030 5099	*0·6*
0·7	+·037 0925^{-}	−·015 0785^{-}	−·036 2959	+·007 8032	+·034 6183	*0·7*
0·8	+·040 0837	+·000 1309	−·036 4311	−·006 9892	+·032 2067	*0·8*
0·9	+·037 1480	+·014 1520	−·030 6973	−·019 4561	+·024 2015^{-}	*0·9*
1·0	+·029 2108	+·024 8341	−·020 5443	−·027 5570	+·012 5634	*1·0*
1·1	+·017 9171	+·030 7754	−·008 0852	−·030 2452	−·000 2002	*1·1*
1·2	+·005 2896	+·031 5046	+·004 3583	−·027 5833	−·011 6000	*1·2*
1·3	−·006 6483	+·027 4720	+·014 7085^{+}	−·020 6206	−·019 6708	*1·3*
1·4	−·016 2101	+·019 8699	+·021 4900	−·011 0828	−·023 3144	*1·4*
1·5	−·022 2864	+·010 3376	+·024 0306	−·000 9592	−·022 4203	*1·5*
1·6	−·024 4563	+·000 6188	+·022 4833	+·007 9130	−·017 7633	*1·6*
1·7	−·022 9630	−·007 7542	+·017 6879	+·014 1799	−·010 7294	*1·7*
1·8	−·018 5798	−·013 6952	+·010 9196	+·017 1592	−·002 9520	*1·8*
1·9	−·012 4032	−·016 6827	+·003 5931	+·016 8681	+·004 0496	*1·9*
2·0	−·005 6239	−·016 7587	−·003 0142	+·013 9076	+·009 1520	*2·0*
2·1	+·000 6813	−·014 4266	−·007 9675^{-}	+·009 2517	+·011 7814	*2·1*
2·2	+·005 7023	−·010 4847	−·010 7807	+·003 9997	+·011 9289	*2·2*
2·3	+·008 9753	−·005 8372	−·011 4193	−·000 8515^{+}	+·010 0479	*2·3*
2·4	+·010 3905^{+}	−·001 3219	−·010 2197	−·004 5721	+·006 8771	*2·4*
2·5	+·010 1377	+·002 4160	−·007 7554	−·006 7797	+·003 2407	*2·5*
2·6	+·008 6140	+·004 9907	−·004 6874	−·007 4374	−·000 1273	*2·6*
2·7	+·006 3171	+·006 2836	−·001 6306	−·006 7851	−·002 7038	*2·7*
2·8	+·003 7444	+·006 4027	+·000 9425^{+}	−·005 2343	−·004 2287	*2·8*
2·9	+·001 3165^{-}	+·005 6125^{-}	+·002 7502	−·003 2536	−·004 6928	*2·9*
3·0	−·000 6697	+·004 2530	+·003 7036	−·001 2712	−·004 2795^{-}	*3·0*
3·1	−·002 0559	+·002 6664	+·003 8746	+·000 3922	−·003 2828	*3·1*
3·2	−·002 8114	+·001 1410	+·003 4425^{+}	+·001 5529	−·002 0245^{+}	*3·2*
3·3	−·003 0043	−·000 1209	+·002 6357	+·002 1606	−·000 7871	*3·3*
3·4	−·002 7645^{-}	−·001 0121	+·001 6797	+·002 2718	+·000 2280	*3·4*
3·5	−·002 2471	−·001 5114	+·000 7608	+·002 0100	+·000 9146	*3·5*
3·6	−·001 6020	−·001 6623	+·000 0056	+·001 5252	+·001 2545^{-}	*3·6*
3·7	−·000 9528	−·001 5475^{+}	−·000 5222	+·000 9601	+·001 2950^{-}	*3·7*
3·8	−·000 3864	−·001 2643	−·000 8138	+·000 4275^{+}	+·001 1208	*3·8*
3·9	+·000 0486	−·000 9048	−·000 9000	+·000 0002	+·000 8276	*3·9*
4·0	+·000 3373	−·000 5432	−·000 8337	−·000 2892	+·000 5010	*4·0*

MODERN PROBLEMS IN VITAL STATISTICS*.

By Professor HARALD WESTERGAARD, Copenhagen.

It is of course impossible in a single lecture to enter into details with regard to all the problems of Vital Statistics. I shall confine myself to some general remarks, giving as I hope a comprehensive view of the subject with some additional observations on certain fields, which—as far as I can see—might be cultivated with good results.

The construction of *life tables* being the first step forward in the evolution of political arithmetic it may be fitting to begin with this problem. In fact the life-table formulae can easily be generalised so as to embrace other problems in vital statistics.

Using the continuous method, which can be traced as far back as Daniel Bernoulli and Duvillard, we can express the number of persons between the ages of x and $x+\epsilon$ (ϵ being infinitely small) at the moment t as $\epsilon p_{x,t}$, p being a function dependent on x and t. If the observations are collected by a life insurance society among assured lives it will present very little difficulty to find the value of $p_{x,t}$ with great accuracy at any given moment. The problem is somewhat more difficult if we have to deal with census results for a whole population, with long intervals—for instance 10 years—between the enumerations. We are then compelled to use more or less refined methods of interpolation. The differences between the results of these various methods are however as a rule not very great, so that we can use any of these methods for all practical purposes without much hesitation. The problem will be to find the variation of the value of $p_{x,t}$ if the time, as well as the age, gets a small increase ϵ, or in other words to find the first differential coefficient:

$$\frac{dp_{x,t}}{dx}+\frac{dp_{x,t}}{dt}.$$

Between the two moments t_2 and t_1 the group in question has embraced the number of individuals

$$\int_{t_1}^{t_2} p_{x,t}\,dt.$$

And dividing this quantity into the number of deaths observed during the same period, viz. ϵd_x, we find the instantaneous rate of mortality (the force of mortality) to be

$$\mu_x=\frac{d_x}{\int_{t_1}^{t_2} p_{x,t}\,dt}.$$

---

* Being a lecture delivered in the University of London to advanced students of statistics.

Having found μ_x we can easily develop various formulae which are useful in mortality statistics. Thus (l_x being the number surviving out of a certain number of individuals of the same age) we shall have:

$$\mu_x = -\frac{1}{l_x}\frac{dl_x}{d_x},$$

or

$$l_{x_2} = l_{x_1} e^{-\int_{x_1}^{x_2} \mu_x d_x},$$

and if μ_x can be considered as constant in the interval x_1 to x_2 we shall have

$$l_{x_2} = l_{x_1} e^{-(x_2-x_1)\mu}.$$

If the interval is unity this equation is reduced to the following:

$$l_{x_1+1} = l_{x_1}\left(1 - \frac{\mu}{1} + \frac{\mu^2}{1.2} - \frac{\mu^3}{1.2.3} + \dots\right).$$

And the probability of a person aged x_1 dying within a year will be:

$$\frac{l_{x_1} - l_{x_1+1}}{l_{x_1}} = \frac{\mu}{1} - \frac{\mu^2}{1.2} + \frac{\mu^3}{1.2.3} - \dots.$$

With approximation the expression on the right side of this equation may be replaced by the following*:

$$\frac{\mu}{1} - \frac{\mu^2}{2} + \frac{\mu^3}{4} - \dots = \frac{\mu}{1+\frac{\mu}{2}},$$

the difference being less than $\frac{\mu^3}{12}$. The probability of dying within unity of time will thus be a trifle smaller than the force of mortality.

Again we can resolve μ into a sum of quantities each one representing the frequency of deaths from various causes, say tuberculosis, alcoholism, etc. If for instance $\mu = \mu' + \mu''$ we can calculate fictitious life tables putting

$$l'_{x_1+1} = l'_{x_1} e^{-\mu'} \text{ and } l''_{x_1+1} = l''_{x_1} e^{-\mu''},$$

and we easily find

$$\frac{l_{x_1+1}}{l_{x_1}} = \frac{l'_{x_1+1}}{l'_{x_1}} \cdot \frac{l''_{x_1+1}}{l''_{x_1}}.$$

If one of the causes of death disappears through some factor of sanitary or social progress (as for instance Jenner's Vaccination against smallpox a century ago) this formula will give us the means to judge of the effect on the life table.

Frequently we desire to measure such influence by means of the *expectation of life*, viz. the average number of years which a person of a given age will attain according to the life table. At the age of x_1, the value of the expectation of life will be

$$E_x = \frac{1}{l_{x_1}}\int_{x_1}^{\omega} l_x dx,$$

where ω is the highest age attainable according to the life table, for instance 100 years.

---

* [Still closer by $\mu/(1+\frac{1}{6}\mu)^3$, the difference being then less than $\frac{1}{216}\mu^4$. Ed.]

If the age-distribution of a population corresponds to that of a life table, as for instance according to Halley's supposition, we can easily prove that the expectation of life is identical with the number of persons out of whom one will die in unity of time. In a moment of length ϵ the number of deaths out of $\int_{x_1}^{\infty} l_x dx$ will be $l_{x_1}\epsilon$, and consequently in unity of time l_{x_1} will die. This problem puzzled our forefathers in the 18th century.

If for instance the rate of mortality of the whole population was 3 per cent. yearly, the expectation of life for a new-born child could be estimated at $33\frac{1}{3}$ years. But at the same time the birth-rate was perhaps 4 per cent., the population thus not being constant; they therefore concluded that the expectation of life would lie somewhere between $33\frac{1}{3}$ and 25 years, for instance:

$$\frac{200}{3+4} = 28\cdot6 \quad \text{or} \quad \tfrac{1}{2}\left(33\tfrac{1}{3} + 25\right) = 29\cdot2.$$

By changing the terms we can easily adapt the preceding formulae to other branches of vital statistics, asking for instance how a group of *bachelors* will be gradually reduced by marriage and mortality, or—in *criminal* statistics—what is the probability that a person of a given age will never be punished for a crime.

Difficulties will rarely meet us with regard to mathematical formulae, the great obstacles as a rule lying in the observations. If these on the whole are at hand they will frequently be more or less defective, and it will be the task of the statistician to find his way in spite of these defects. Such difficulties for instance appear in *marriage* statistics, where the observations frequently are inaccurate because the weddings registered do not correspond exactly to the population enumerated within the district. It is relatively easy to get tolerably good observations on *deaths*, but it is otherwise with regard to *invalidity* and *infirmity*. We can calculate the surplus *emigration* from a country, adding perhaps some partial observations on ages of emigrants to America or Australasia, but it is most difficult to get exhaustive statistical observations on *migrations* within the country.

Whatever subject we have to deal with in the first instance the problem will be to find the value of a fraction, i.e. the rate of frequency of a certain event.

The problem can easily be solved if we have sufficiently accurate observations both on the numerator and the denominator of the fraction. But frequently the *denominator* is unknown. We are perhaps possessed of accurate accounts of the distribution of births, deaths and marriages in the single months of the calendar year, whereas the corresponding group of the population is unknown. Still we are not altogether excluded from drawing conclusions if we feel justified in supposing that the population concerned does not vary too much within the space of time concerned. As a rule we may without much hesitation make investigations as to the influence of season *in vital statistics* without taking the population into consideration or without noticing the secular trend of the movements of mortality or fertility; on the other hand in *Economic Statistics* we must be much more careful, the movements being so conspicuous that we cannot compare January and December in the same calendar year without taking these movements into consideration. As problems of this kind in vital statistics I may mention the prevalence of certain diseases in months which have particularly high rates of infant mortality. Without knowing the number of infants exposed to risk we may ask whether an excessive heat in August or September will prove fatal to babies under 1 year, etc.

It is not uncommon to meet proposals for dealing with one-sided material, as for instance the distribution of deaths according to causes. The well known Hungarian statistician Kőrösy proposed for instance to compare the risk of dying from certain causes of death in various classes of society by calculating the ratio of the number of these cases to the total number of deaths in the class concerned. If in one class of society tuberculosis caused 30 deaths out of 100, in another one only 20, then he felt justified in concluding that tuberculosis was more dangerous in the former group than in the latter. Kőrösy was well aware that the rate of mortality from all causes might be as 2 to 3 in the two classes of society; this being the case, the frequency of death from tuberculosis would be exactly the same. Possibly even the real death rate from tuberculosis might be smaller in the former class than in the latter. His idea was however that if the rates of mortality from all causes were as 2 to 3, then normally the same ratio ought to hold for the various causes of death, and any excess above what he thus considered as normal should be interpreted as testifying to an increased influence of the disease concerned.

Corresponding proposals have been recently made with the view to constructing mortality tables from observations on causes of death upon the supposition that we know the frequency distribution according to some formula. But even if a mortality table calculated in this way may seem remarkably trustworthy we have evidently no security that the rates of mortality which we find provide the true values. These defective observations cannot possibly replace complete observations on the denominator as well as on the numerator.

Turning to the *opposite* case: the denominator being known, whereas we have no observations on the numerator, we shall find somewhat better chances of reaching tolerably safe conclusions. In fact the census gives us the results of the movements of the population. Knowing for instance the distribution of the population above a certain age between bachelors, widowers, etc. we have some evidence with regard to the chances of entering into married life. An example may illustrate this side of the problem.

Let us from the Census calculate the *relative* numbers of bachelors at each age. Thus according to the Danish Census 1911 among 1000 males aged 40—45, 115 were still bachelors. In the following quinquennial age the quota was reduced to 96, and in the ages 50—55, 85 out of 1000 were still unmarried. The question is whether we can use such numbers to find the rate of marriage of bachelors at various ages.

Let the number of bachelors aged x to $x+\epsilon$ be $\epsilon b_{x,t}$ and let the rate of marriage be m_x, whereas δ_x' signifies the rate of decrease through mortality and migration of bachelors. Then we shall have as the result of the changes through the small element of time:

$$\frac{db_{x,t}}{dx}+\frac{db_{x,t}}{dt}=-b_{x,t}(\delta_x'+m_x),$$

whereas for the whole population we have:

$$\frac{dp_{x,t}}{dx}+\frac{dp_{x,t}}{dt}=-p_{x,t}\,\delta_x,$$

where δ in the same way represents the decrease on account of mortality and migration.

Lastly we can take the ratio $r_{x,t}=\frac{b_{x,t}}{p_{x,t}}$ and we shall find by differentiation:

$$\frac{dr_{x,t}}{dx}+\frac{dr_{x,t}}{dp}=-r_{x,t}(\delta_x'-\delta_x+m_x).$$

Generally it will not be difficult to calculate the values on the left side of the equation with sufficient approximation. Knowing, for instance, the relative numbers at two consecutive censuses with an interval, say, 10 years, it is a matter of interpolation to find the variation according to time and age. Thus the rate in 1901 at the ages 20—25 being 881 and 10 years later 876 we may estimate the yearly rate of decrease according to time at 0·0006.

It is a well known fact that mortality among bachelors is higher than in the remaining population. Probably also migrations are more prevalent. Thus we have $\delta_x'>\delta_x$. If we therefore leave $\delta_x'-\delta_x$ out of consideration, thus putting

$$\frac{dr_{x,t}}{dx}+\frac{dr_{x,t}}{dt}=-r_{x,t}\,m_x,$$

we shall find a value of m_x which is somewhat too high. But the difference between this value and the true one will as a rule be comparatively small. In the period of life in which most marriages take place the rate of marriage will be 10 per cent. or more, whereas the force of mortality and probably often the rate of migration will be only a few per mille. This being the case the difference $\delta_x'-\delta_x$ cannot count for much, so that the values calculated from the equation above will on the whole be fairly correct. In fact the results are in good harmony with the values found directly by comparing the numbers of marriages to the enumerated numbers of bachelors.

It would of course be superfluous to make calculations of this kind if we can get direct observations. But the method may be of use in *retrospective* calculations concerning census reports for remote periods. Thus we are possessed of a highly interesting Danish Census for 1787 giving the structure of the society in those times—much more simple indeed than now-a-days. Using the method which I have described we find the rate of marriage for males aged 25 to be only 4 per cent. whereas it was about 12 per cent. in 1911—20. But at 35 it was in 1787 14 per cent. against 10 now-a-days, and at 45 we find 13 per cent. against only 3 per cent. at present. These results show a striking contrast to modern experience.

Another curious problem is presented by the *religious* statistics. In Denmark every census gives observations on the various confessions. An increasing number of both sexes declare themselves as not belonging to any church. Taking 1901 as basis we may ask how many persons should be expected to be standing outside any confession 10 years later in 1911. Let the rate of withdrawals over and above the rate of re-entrance be $\omega_{x,t}$, then in a given element of time ϵ we shall have $\omega_{x,t}\epsilon$ surplus withdrawals. If the number of persons at the first census be $p_{x,t}$, we shall have as the corresponding number 10 years later $p_{x+10,t+10}$ and after n years $p_{x+n,t+n}$. The number of withdrawals between $t+n$ and $t+n+\epsilon$ will be:

$$\omega_{x+n,\,t+n}\cdot p_{x+n,\,t+n}\,\epsilon,$$

and that number will probably be reduced to

$$\omega_{x+n,\,t+n}\cdot p_{x+10,\,t+10}\,\epsilon$$

at the following census. For the whole interval there will thus be a gain of

$$p_{x+10,\,t+10}\int_0^{10}\omega_{x+n,\,t+n}\,dn,$$

or approximately:

$$10p_{x+10,\,t+10}\,\omega_{x+5,\,t+5}.$$

The number of males belonging to some confession at the ages of 25—30 was in 1911 about 98,000. The surplus of persons not belonging to a confession was 1050. The rate of withdrawals for the ages 20—25 will thus probably be 0·0011. For 30—35 we find 0·0006. Above 40 the rate of withdrawals is reduced to a very small value.

Just the same method may be applied to the statistics of *migrations*. Suppose as usual

$$\delta_x=\mu_x+i_x,$$

where μ_x is the force of mortality and i_x the surplus rate of migration (supposed constant with regard to t within the interval), then we shall have:

$$p_{x+10,\,t+10}=p_{x,\,t}\cdot e^{-\int_0^{10}\delta_{x+n}dn},$$

or approximately

$$e^{-10\delta_{x+5}}=\frac{p_{x+10,\,t+10}}{p_{x,\,t}}.$$

This equation will give us δ_x and subtracting μ_x we find i_x.

The main results for Denmark in respect to males are the following for the years 1901—11:

Age	Force of mortality p_m	Rate of net-emigration p_m
5—10 years	3	3
10—15 ,,	2	5
15—20 ,,	3	15
20—25 ,,	4	15
25—30 ,,	5	4
30—35 ,,	5	1

After 35 the rate of emigration is very small. It is principally the young men who leave the country. As to the females emigration is much smaller and the influence of age is less conspicuous.

These results can again be used in other domains where direct observations are lacking and where it will therefore be reasonable to try indirect methods. Let us endeavour to find probable rates of mortality for the *feeble-minded*, using the Danish Census of 1911 as base. As usual we have to find the value of

$$r_{x,t}(\mu_x' + i_x' - \mu_x - i_x - m_x),$$

where m_x is the rate of fresh cases of feeble-mindedness. This quantity seems however to be very small. Most of the persons registered as feeble-minded seem to be born as such. We can thus leave m_x out of consideration, the result probably being that the rate of mortality of the feeble-minded is underestimated a little. We can also probably consider i_x', the rate of emigration of feeble-minded, as a negligible quantity so that we have only to deal with

$$\mu_x' - \mu_x - i_x.$$

The results will be the following rates of mortality for male idiots:

Age	Idiots p_m	Normal persons p_m
20—25	30	4
25—30	24	4
30—35	26	5
35—40	28	6
40—45	30	7
45—50	36	10
50—55	40	13
55—60	40	19
60—65	59	27

These minimum values are in good harmony with an investigation which I made several years ago concerning mortality in asylums for the feeble-minded, basing my conclusions on direct observations on deaths and on numbers exposed to risk.

The table testifies to a very high mortality among the feeble-minded especially in earlier years of life so that the rates of mortality of the feeble-minded remain nearly constant through a long period of life.

Studying statistical observations on other defects we find curious results with regard to *blindness*, but here it is the rate of the incidence of blindness and not the mortality of the blind which we are approximating to. Supposing that few blind persons recover and that the mortality of blind persons will differ little from that of the general population, and confining ourselves only to the advanced years of age where migrations are trifling, we can try to estimate the chances of becoming blind in various ages, the expression $\mu_x' + i_x' - \mu_x - i_x - m_x$, where m_x is the rate of becoming blind, being reduced to $-m_x$. We conclude from the observations that out of 100,000 males at the age of 55–60 only four will become blind yearly, in the following quinquennium of age 6, etc.

For sociological purposes however it may often be considered sufficient to use still simpler methods to ascertain the influence of migrations and other displacements of population. We may confine ourselves to finding the gain or loss from one census to another.

Taking as instance the numbers of *Roman Catholics* in Denmark, 1901 and 1911, we find that 623 females aged 5–15 years belonged to this church in 1901. According to the age-distribution 10 years later we should expect to find 591 in 1911 in the age group 15–25. The number observed was however 1265 so that the gain was 674.

The above quoted census of 1787 gives interesting evidence as to the *displacements* between various classes of society, in the younger years for instance from the group of masters to that of journeymen, in mature age in the opposite direction. In the rural districts the class of servants gains large numbers from the peasants, the small-holders and others, till in the age group 30–40 the tide is turning. As one of the burning questions of the day which may be treated by this elementary method I can mention *migrations* within a country from the *rural districts* to the *towns*.

In conclusion I may refer to the complicated problem of *birth statistics*. First of all we have to find the birth-rates under various circumstances, for instance according to conjugal condition, age of mother, etc. If sufficient observations are at hand we can calculate tables corresponding to mortality tables, comparing the results for various epochs in order to follow the decrease of fertility which has been so conspicuous in the last decennia. Observations have been obtained in various countries, in Scandinavia, New Zealand and Australia. In France we can follow the movement back until 1896, in Denmark even as far as 1870. Behind these problems other important questions arise, for instance to find the probability of a wife at a certain age bearing a child when we take into consideration the duration of her marriage and the number of children which she has borne previously. Further the influence of the size of the family or of the rapidity with which births follow each other, on the infant mortality. Several of these problems may be treated with advantage by means of privately collected observations such as Ch. Ansell's Family Statistics (1874). In North America useful observations of this kind have been collected of late. But on the whole the field is not yet exhausted.

If we use Census results only we shall have difficulty in applying the indirect methods of finding the birth-rates just described, the decrease of fertility complicating the problem. But it will not be difficult to draw indirect conclusions of a more elementary and preliminary character from Census reports. A Danish Census of 1880 in Copenhagen provided material of this kind with particulars as to the number of children alive or dead at the moment of the Census. These observations gave evidence of a considerable difference between various classes of society, and just the same was the case in a later Census for the whole country. Important conclusions could be drawn concerning the vitality of the children born in families of various size testifying to an enormous harvest of death in marriages with a very numerous offspring. It is unnecessary to add that investigations of this kind will necessitate some standardization of the observations in order to outbalance the effect of disturbing causes, as for instance of different age-distributions in various classes of society. Similar observations were made later on in other countries; thus in England by the Census of 1911, Professor Pearson having previously drawn attention to these problems. In Norway the Census of 1920 gave observations of a similar kind. It is to be hoped that several countries will follow Norway and that on the whole birth statistics will be recognized as one of the most fertile subjects for both modern statistics and sociology.

Generally I hope it will flow from what I have said that our statisticians of to-day need not fear that they will be thrown out of work. On the contrary they are likely to suffer from a cumbersome *embarras de richesses.*

ON THE EXTREME INDIVIDUALS AND THE RANGE OF SAMPLES TAKEN FROM A NORMAL POPULATION.

By L. H. C. TIPPETT, B.Sc. Lond.

THE problem of the range of samples arises as a special case of Galton's Difference Problem, first given by Professor K. Pearson in 1902 (1). Together with the allied problem of the extreme individuals, it has engaged the attention of other writers (2) (3) (4), but a complete solution has not yet been given. This would involve the determination of the distribution of the range and of the extremes for a large number of samples. Attempts are here made in some measure to supply this deficiency.

I. THE FIRST OR LARGEST INDIVIDUAL.

Let us consider samples of size n to be taken from an infinite population represented by the curve $y = \phi(x)$, which we will suppose extends to infinity in each direction, and let

$$\alpha_p{}^* = \int_{-\infty}^{x_p} \phi(x)\,dx, \quad \text{where} \quad \int_{-\infty}^{+\infty} \phi(x)\,dx = 1.$$

Then the chance that the n individuals of the sample shall be smaller than x_p equals $\alpha_p{}^n$, and this is the chance that the largest shall lie between $x = -\infty$ and $x = x_p$.

Hence, if $y = f(x)$ be the distribution of the largest individual in samples of n $\left(\text{where} \int_{-\infty}^{+\infty} f(x)\,dx = 1\right)$, we have for the probability integral of this curve

$$\int_{-\infty}^{x_p} f(x)\,dx = \alpha_p{}^n \quad \text{.................................(1).}$$

This has been determined for samples from a normal population for several values of n, and some of the results are given in Table IX at the end of the paper†. The values of the variate are all in terms of the standard deviation of the original population. The constants given in Table I (the mean, standard deviation, β_1 and β_2) have been found from the grouped form of the distribution, using Sheppard's corrections, and are also plotted in Diagrams I, II and III. The mean is half the mean range, which is given in Diagram V. Since the normal curve is symmetrical, the distribution of the smallest individual is quite similar, excepting that the variate has the opposite sign.

---

* This α must not be confused with the α of Dr Sheppard's Tables of the Probability Integral; throughout this paper, the α used is equivalent to $\frac{1}{2}(1+\alpha)$ of those tables.

† Full tables are in manuscript, and it is hoped that they will be published later.

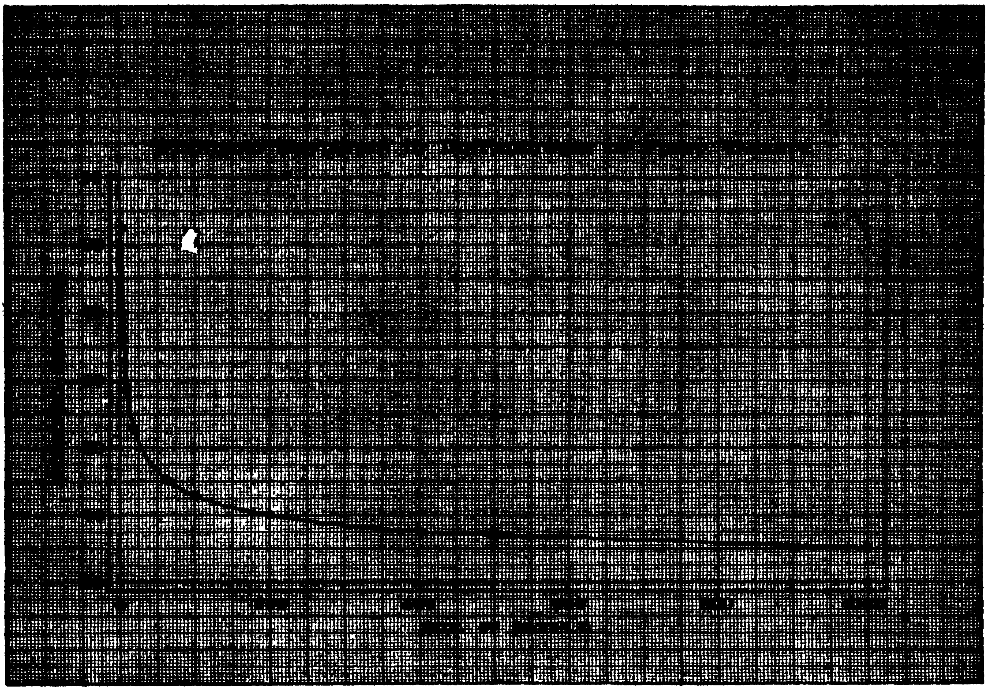


Diagram I.

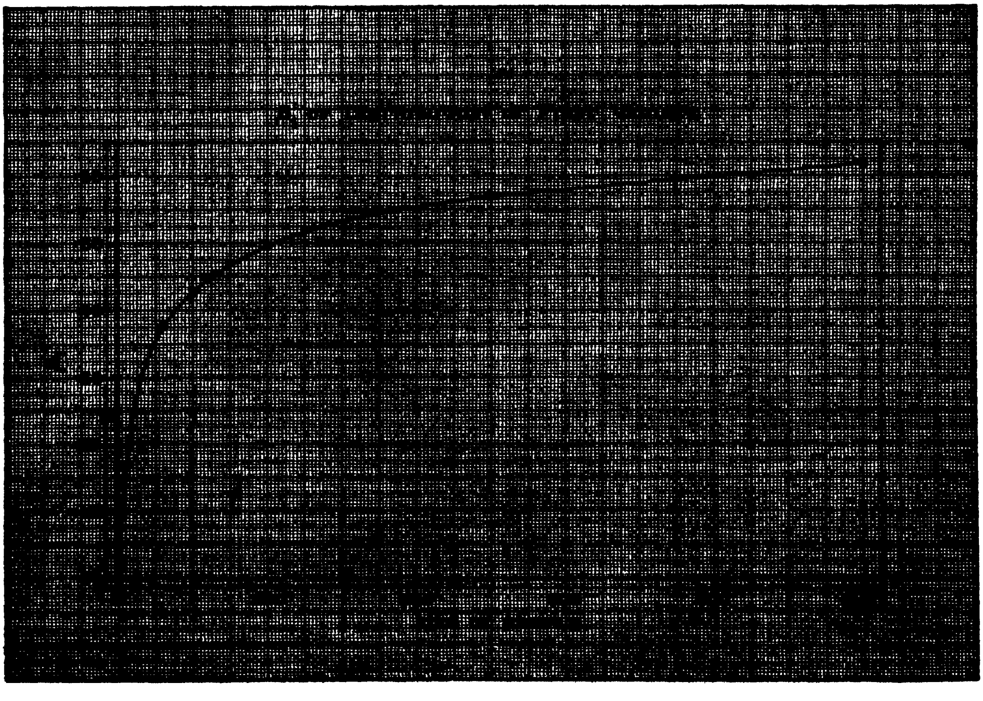


Diagram II.

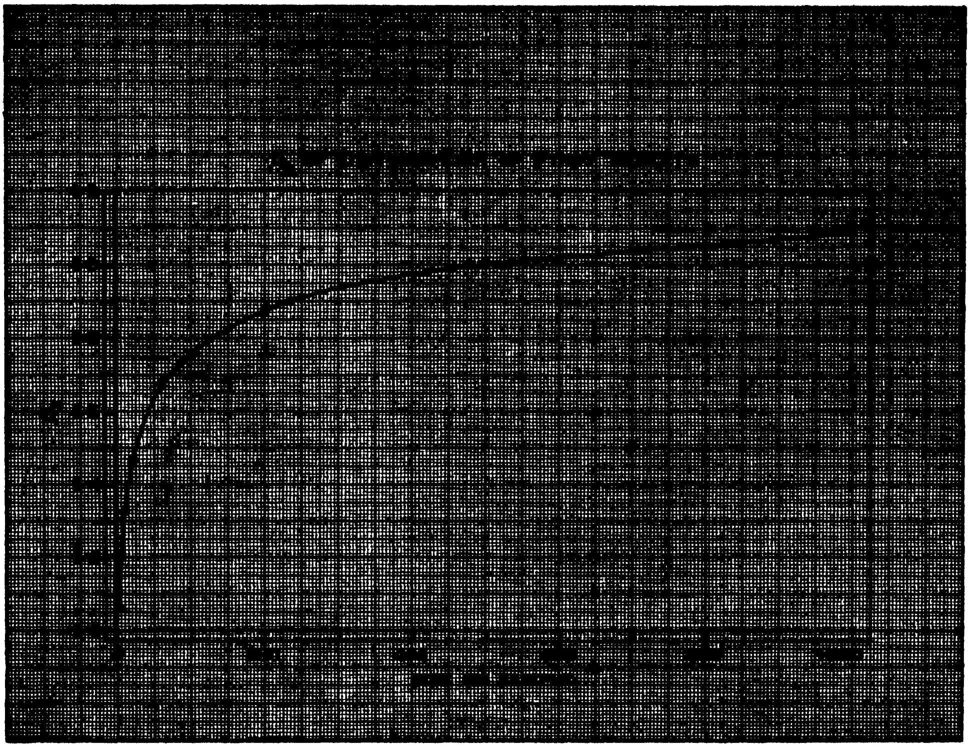


Diagram III.

TABLE I.

Constants of the Distribution of the Largest Individual in Samples from a Normal Population.

Size of Sample n	Mean	Standard Deviation	β_1	β_2
2	·56419	·8257	·019	3·062
5	1·16297	·6690	·092	3·202
10	1·53875	·5868	·168	3·331
20	1·86747	·5251	·251	3·469
60	2·31928	·4545	·376	3·677
100	2·50759	·4294	·429	3·765
200	2·74604	·4009	·495	3·875
500	3·03670	·3704	·570	4·003
1000	3·24144	·3514	·618	4·088

It is interesting to note that we have here an example of a statistical variate, the distribution of which diverges more from normality as the sample from which it is measured increases in size, at least this is so to the limit $n = 1000$.

The tables of the distribution of the extremes may well be used for deciding whether or not any outlying member of a sample should be rejected. The deviation of this individual from the mean is measured in terms of the standard deviation, and from Table IX, the chance of a deviation occurring as great or greater than this is found. This chance is $(1 - \alpha_E)$, where α_E is the value read directly from the table, and if it falls below a certain limiting value, the individual is rejected. In Table II are given values of the deviations of the extreme variate corresponding to limiting chances of 1 % and 5 %. These are plotted on Diagram IV, which can be used when the samples are of intermediate size.

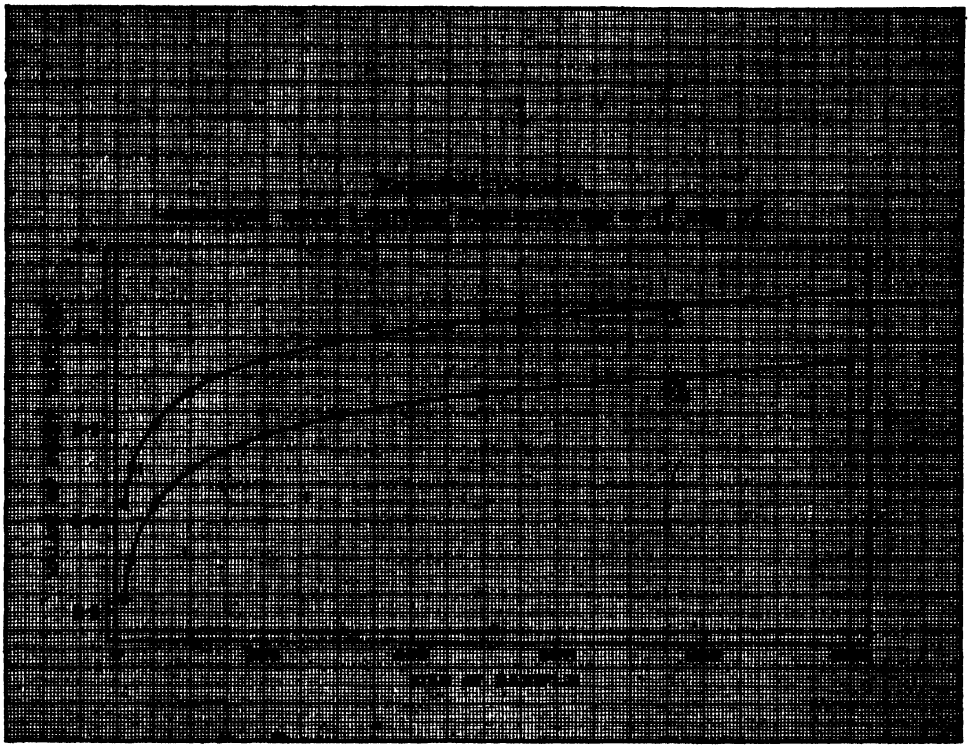


Diagram IV.

TABLE II.

Values of the Extreme Variate occurring with Limiting Probabilities of 5 °/ₒ and 1 °/ₒ in Samples of Size n.

n	5 per cent.	1 per cent.
10	2·5695	3·0924
20	2·7992	3·2893
50	3·0831	3·5393
100	3·2838	3·7180
200	3·4740	3·8894
500	3·7126	4·1063
1000	3·8845	4·2638

II. On the Theoretical Range.

The range is defined as the difference in character between the largest and smallest individuals of a sample. It has not been found possible to write down the distribution of this in any useful form, so the procedure has been to find those constants involving the first four moments, so that an appropriate Pearson Curve can be fitted. It is usually found that curves obtained in this way give good fits to actual data, so the method will be adequate for practical purposes.

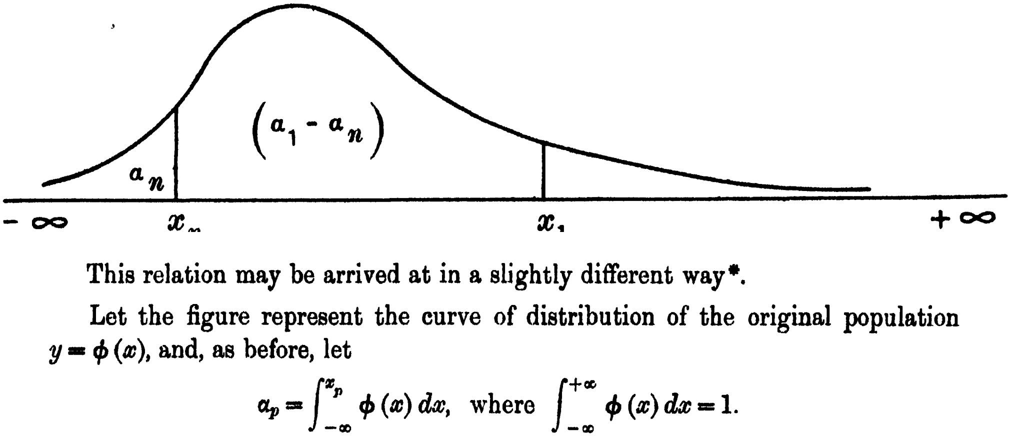
(i) *The Mean.*

Professor Pearson in his paper (1) gave an expression for the mean difference between the pth and $(p+1)$th individuals

$$\chi_p = \frac{n!}{(n-p)!\,p!}\int_{-\infty}^{+\infty} \alpha^{n-p}(1-\alpha)^p\,dx \text{ .... (2)},$$

from which, by summing for all values of p from 1 to $n-1$, we obtain for the mean range

$$w = \int_{-\infty}^{+\infty} \{1-(1-\alpha)^n - \alpha^n\}\,dx \text{ .... (3)}$$



This relation may be arrived at in a slightly different way*.

Let the figure represent the curve of distribution of the original population $y = \phi(x)$, and, as before, let

$$\alpha_p = \int_{-\infty}^{x_p} \phi(x)\,dx, \quad \text{where} \quad \int_{-\infty}^{+\infty} \phi(x)\,dx = 1.$$

Let x_1 be the character of the first individual, and x_n that of the last in a sample of size n.

Then, if we suppose the original population to be infinite, the chance that we have one individual at x_1, one at x_n and $n-2$ between

$$= \frac{n!}{(n-2)!}(\alpha_1 - \alpha_n)^{n-2}\,d\alpha_1\,d\alpha_n \text{ .... (4)}.$$

Whence we have for the mean range

$$\overline{w} = \frac{n!}{(n-2)!}\int_{x_1=-\infty}^{x_1=+\infty}\int_{x_n=-\infty}^{x_n=x_1} (\alpha_1-\alpha_n)^{n-2}(x_1-x_n)\,d\alpha_1\,d\alpha_n \text{ .... (5)}.$$

But $$(\alpha_1-\alpha_n)^{n-2} = \underset{s=0}{\overset{s=n-2}{S}} (-1)^s \frac{(n-2)!}{s!\,(n-2-s)!}\,\alpha_1^{\,n-2-s}\,\alpha_n^{\,s}.$$

---

* This proof was suggested by Prof. Pearson. Mr J. O. Irwin gives yet another proof in (5).

Therefore

$$\bar{w} = n! \mathop{S}_{s=0}^{s=n-2} \frac{(-1)^s}{s!\,(n-2-s)!} \int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-2-s} . \,d\alpha_1 \int_{x_n=-\infty}^{x_n=x_1} \alpha_n^s (x_1 - x_n)\, d\alpha_n \ldots(6).$$

Now

$$\int_{x_n=-\infty}^{x_n=x_1} \alpha_n^s (x_1 - x_n)\, d\alpha_n = \left[\frac{(x_1 - x_n)\, \alpha_n^{s+1}}{s+1}\right]_{x_n=-\infty}^{x_n=x_1} + \frac{1}{s+1} \int_{-\infty}^{x_1} \alpha_n^{s+1}\, dx_n.$$

The term in square brackets vanishes at both limits, so that we obtain, by substituting in (6),

$$\bar{w} = n! \mathop{S}_{s=0}^{s=n-2} \frac{(-1)^s}{(s+1)!\,(n-2-s)!} \int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-2-s}\, U d\alpha_1,$$

where

$$U = \int_{-\infty}^{x_1} \alpha_n^{s+1}\, dx_n.$$

Let

$$\theta = \int_{-\infty}^{1} \alpha_1^{n-2-s}\, d\alpha_1 = \frac{1 - \alpha_1^{n-1-s}}{n-1-s}.$$

Then,

$$\bar{w} = -n! \mathop{S}_{s=0}^{s=n-2} \frac{(-1)^s}{(s+1)!\,(n-2-s)!} \int_{x_1=-\infty}^{x_1=+\infty} U \frac{d\theta}{d\alpha_1} d\alpha_1 \ldots\ldots\ldots\ldots(7).$$

Integrating by parts, we obtain

$$\bar{w} = n! \mathop{S}_{s=0}^{s=n-2} \frac{(-1)^s}{(s+1)!\,(n-2-s)!} \left\{- \left[U\theta\right]_{x_1=-\infty}^{x_1=+\infty} + \int_{-\infty}^{+\infty} \theta \frac{dU}{dx_1}\, dx_1\right\}.$$

Again the term in square brackets vanishes, and

$$\frac{dU}{dx_1} = \alpha_1^{s+1},$$

so that we have

$$\mathop{S}_{s=0}^{s=n-2} (-1)^s \frac{n!}{(s+1)!\,(n-1-s)!} \int_{-\infty}^{+\infty} (1 - \alpha_1^{n-1-s})\, \alpha_1^{s+1}\, dx,$$

leading to the result of equation (3),

$$\bar{w} = \int_{-\infty}^{+\infty} \{1 - (1-\alpha)^n - \alpha^n\}\, dx.$$

(ii) *The Higher Moments.*

In a similar manner, it is possible to obtain an expression for the higher moments.

From (6) we see that the mean value of the mth moment (taken about the mean)

$$\mu_m = n! \mathop{S}_{s=0}^{s=n-2} \frac{(-1)^s}{s!\,(n-2-s)!} \int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-2-s}\, d\alpha_1 \int_{x_n=-\infty}^{x_n=x_1} \alpha_n^s (x_1 - x_n - \bar{w})^m d\alpha_n \ldots(8).$$

Now

$$\int_{x_n=-\infty}^{x_n=x_1} \alpha_n^s (x_1 - x_n - \bar{w})^m\, d\alpha_n = \left[\frac{(x_1 - x_n - \bar{w})^m \alpha_n^{s+1}}{s+1}\right]_{x_n=-\infty}^{x_n=x_1}$$

$$+ \frac{m}{s+1} \int_{x_n=-\infty}^{x_n=x_1} (x_1 - x_n - \bar{w})^{m-1} \alpha_n^{s+1} dx_n = \frac{(-\bar{w})^m \alpha_1^{s+1}}{s+1} + \frac{m}{s+1} V \text{ (say)}.$$

On substituting in (8), we obtain

$$\mu_m = \overset{s=n-2}{\underset{s=0}{S}} (-1)^s \frac{n!}{(s+1)!\,(n-2-s)!} \left\{ (-\overline{w})^m \int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-1}\, d\alpha_1 + m \int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-2-s}\, V d\alpha_1 \right\}.$$

Let $I = \int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-2-s}\, V d\alpha_1$, and as before, let $\theta = \int_{\alpha_1}^{1} \alpha_1^{n-2-s}\, d\alpha_1 = \frac{1-\alpha_1^{n-1-s}}{n-1-s}$.

Then $$I = -\int_{x_1=-\infty}^{x_1=+\infty} \frac{d\theta}{d\alpha_1} V d\alpha_1 = -\left[\theta V\right]_{x_1=-\infty}^{x_1=+\infty} + \int_{-\infty}^{+\infty} \theta \frac{dV}{dx_1}\, dx_1.$$

The term in square brackets vanishes at both limits, and

$$\frac{dV}{dx_1} = (-\overline{w})^{m-1} \alpha_1^{s+1} + (m-1) \int_{-\infty}^{x_1} (x_1 - x_n - w)^{m-2} \alpha_n^{s+1}\, dx_n,$$

so that $$I = (-w)^{m-1} \int_{-\infty}^{+\infty} \frac{(1-\alpha_1^{n-s-1}).\,\alpha_1^{s+1}}{n-s-1}\, dx_1 + \frac{m-1}{n-s-1} \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} (1-\alpha_1^{n-s-1})\, \alpha_n^{s+1} (x_1 - x_n - \overline{w})^{m-2}\, dx_1 dx_n,$$

and $$\int_{x_1=-\infty}^{x_1=+\infty} \alpha_1^{n-1}\, d\alpha_1 = \frac{1}{n}.$$

Substituting in (8), we obtain

$$\begin{aligned}\mu_m = {} & \overset{s=n-2}{\underset{s=0}{S}} (-1)^s \frac{(n-1)!}{(s+1)!\,(n-2-s)!} (-\overline{w})^m \\ & + (-\overline{w})^{m-1} \overset{s=n-2}{\underset{s=0}{S}} (-1)^s \frac{n!}{(s+1)!\,(n-s-1)!} m \int_{-\infty}^{+\infty} (1-\alpha_1^{n-s-1})\, \alpha_1^{s+1}\, dx_1 \\ & + m(m-1) \overset{s=n-2}{\underset{s=0}{S}} (-1)^s \frac{n!}{(s+1)!\,(n-s-1)!} \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} (1-\alpha_1^{n-s-1})\, \alpha_n^{s+1} \\ & \qquad \times (x_1 - x_n - w)^{m-2}\, dx_1 dx_n.\end{aligned}$$

If n is even, the first term reduces to $(-\overline{w})^m$.

The second term $= m\overline{w}(-\overline{w})^{m-1} = -m(-\overline{w})^m$.

The third term becomes

$$m(m-1) \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} \{1 - \alpha_1^n - (1-\alpha_n)^n + (\alpha_1 - \alpha_n)^n\} (x_1 - x_n - \overline{w})^{m-2}\, dx_1 dx_n.$$

Hence, for even values of n, we have

$$\mu_m = m(m-1) \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} \{1 - \alpha_1^n - (1-\alpha_n)^n + (\alpha_1 - \alpha_n)^n\} (x_1 - x_n - \overline{w})^{m-2}\, dx_1 dx_n - (m-1)(-\overline{w})^m \quad \text{......(9)},$$

and, on putting $m = 2$,

$$\mu_2 = \sigma^2 = 2 \int_{-\infty}^{+\infty} \int_{-\infty}^{x_1} \{1 - \alpha_1^n - (1-\alpha_n)^n - (\alpha_1 - \alpha_n)^n\}\, dx_1 dx_n - \overline{w}^2 \quad \text{............(10)}.$$

Similarly by putting $m = 3$ and $m = 4$, the formulae for the third and fourth moments can be written down.

III. ON THE RANGE (RESULTS FOR A NORMAL CURVE).

(i) *The Mean.*

The mean range for a normal distribution has been found for all samples from two to one thousand, and is tabled at the end of the paper. A framework of values was found by direct computation of equation (3), using quadratures, and this was filled in by interpolation, using first Lagrangian Formulae, and finally a difference formula. The accuracy of the quadrature was checked by seeing that the full number of ordinates gave sensibly the same result as half the number. Here again, as indeed throughout the paper, the values are for a population having a unit standard deviation, and in any given case must be multiplied by the actual value of this quantity to obtain the absolute range. Diagram V illustrates the results graphically.

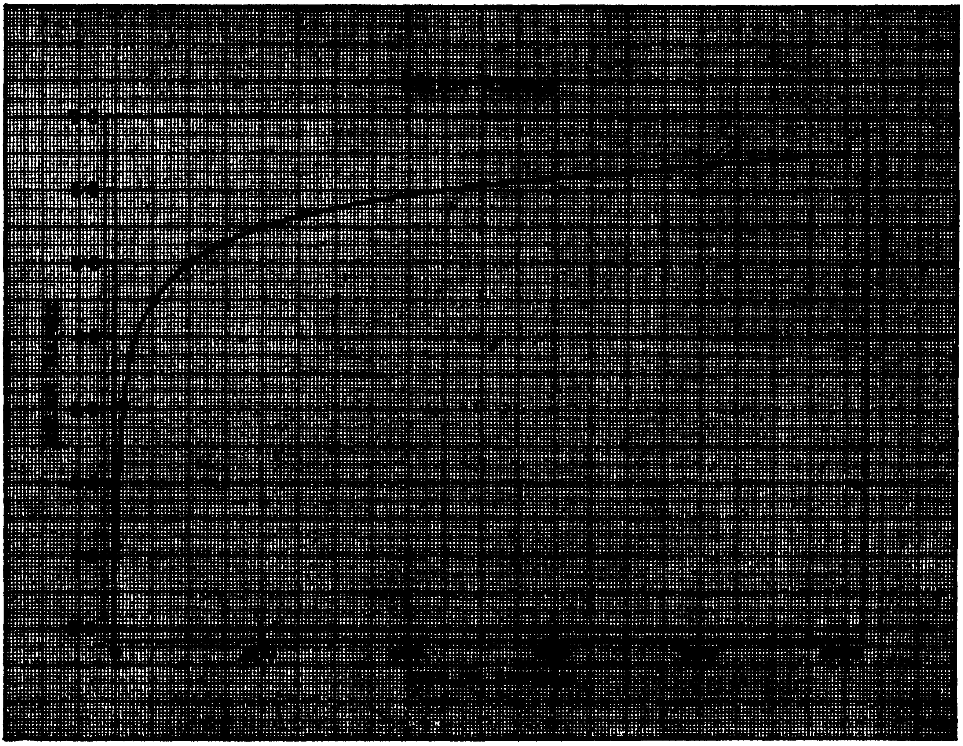


Diagram V.

The determinations of Bortkiewicz (2) and Dodd (3) are not always in good agreement with those of this paper; but they are admittedly approximations. By one method however, Bortkiewicz does obtain good values, as may be seen from Table III.

TABLE III.

Comparison of Values of Mean Range.

Size of Sample	Bortkiewicz	True Mean
8	2·83	2·8472
24	3·90	3·8953
48	4·48	4·4662
100	5·03	5·0152
450	6·01	6·0090
1000	6·48	6·4829

(ii) *Higher Moments.*

The second moment, and hence the standard deviation of the distribution, was determined in several cases by means of equation (10). The work is very laborious, as it involves cubature, and even so, the result can only be given to a few figures. It is believed that the values given in Table IV are correct to the last figure. From the curve of Diagram VI, the standard deviation can be read off for samples of intermediate size.

Much time was spent in trying to evaluate the third and fourth moments from equation (9) by the same method, but there are many difficulties, and the results

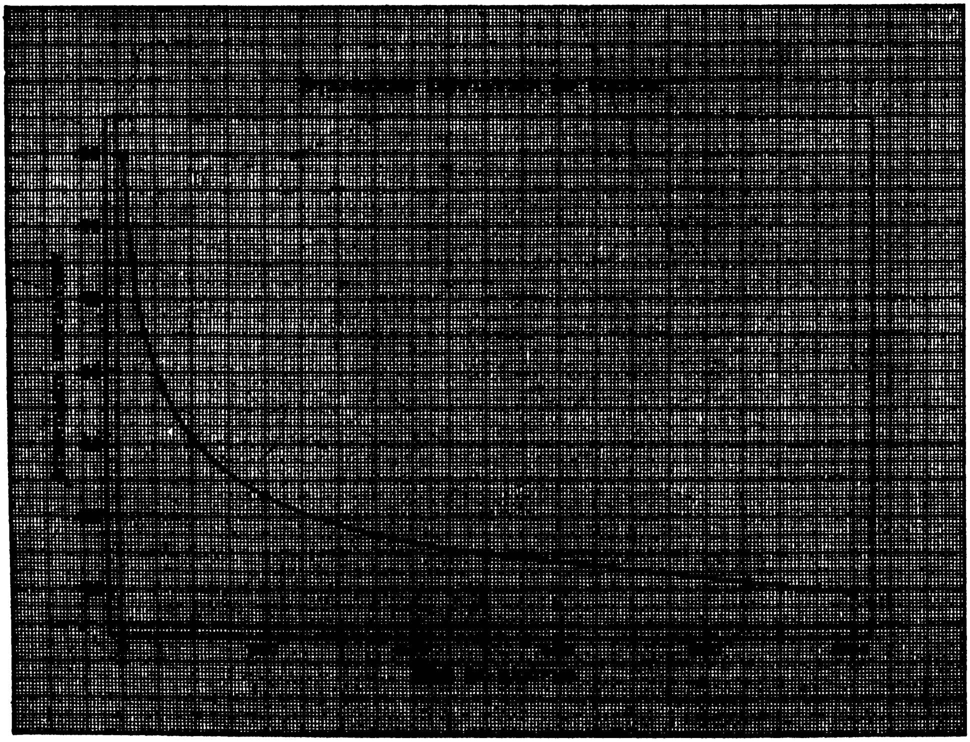


Diagram VI.

were most irregular. It will be seen that the formula (9) consists of two parts, which are subtracted from one another. These are nearly equal, so that they must separately be determined very accurately if their difference (giving the required moment) is to be relied upon. For example, in determining the fourth moment for $n = 200$, it was found that the term integrated was about 2729·8342, and the other term ($3\bar{w}^4$) 2729·4203, giving the fourth moment equal to ·4140. In order that this should be correct to three figures, the integrated term would have to be found correct to seven figures, an accuracy which is very difficult to obtain by any cubature formula with a reasonable expenditure of labour. In this case about 1500 ordinates were calculated, and when they were summed twice by means of the Euler-Maclaurin quadrature formula, the value 2729·8342 was obtained, while Weddle's Rule gave 2729·7157 for the integrated term. These two give widely different, and hence wholly unreliable, values for the fourth moment. The discrepancies are accounted for by the fact that insufficient ordinates were taken; as far as the number of figures used is concerned, the result should be correct to three decimal places. The difficulty was more acute for the larger samples, and for the higher moments, but in all cases it prevented the determination of reliable values of μ_3 and μ_4*. Consequently, a method of obtaining these constants from those of the separate distributions of the first and last individuals was resorted to.

Let us consider a system of variates u, v and w, measured from their respective means, and let $w = u - v$. Then, quite generally, we have

$$\bar{w} = \bar{u} - \bar{v} = 0,$$

$$\left.\begin{aligned} {}_w\mu_2 &= {}_u\mu_2 - 2r\sqrt{{}_u\mu_2 \cdot {}_v\mu_2} + {}_v\mu_2 \\ {}_w\mu_3 &= {}_u\mu_3 - 3p_{21} + 3p_{12} - {}_v\mu_3 \\ {}_w\mu_4 &= {}_u\mu_4 - 4p_{31} + 6p_{22} - 4p_{13} + {}_v\mu_4 \end{aligned}\right\} \quad \text{.....................(11)},$$

where p_{rs} is the mean value of the product $u^r v^s$. Now, if w be the range, and u and v be the first and last individuals respectively, of a sample, then

$$\begin{aligned} \bar{u} &= -\bar{v}, & {}_u\mu_2 &= {}_v\mu_2, \\ p_{12} &= -p_{21}, & {}_u\mu_3 &= -{}_v\mu_3, \\ p_{31} &= p_{13}, & {}_u\mu_4 &= {}_v\mu_4, \end{aligned}$$

since the samples are taken from a distribution symmetrical about its origin, and we have for the moments of the distribution of the range in terms of those of the largest variate

$$\begin{aligned} \sigma_w{}^2 = {}_w\mu_2 &= 2{}_u\mu_2(1 - r) \\ {}_w\mu_3 &= 2{}_u\mu_3 - 6p_{21} \\ {}_w\mu_4 &= 2{}_u\mu_4 - 8p_{31} + \end{aligned} \quad \text{.(11)}\mathit{bis}.$$

---

* As an example, some of the actual values of the β's obtained by cubature are here given:

n	*10*	*20*	*60*	*100*	*200*	*500*
β_1	·152	·153	·214	·209	·255	·265
β_2	3·37	3·44	3·34	3·71	4·02	3·81

Moreover, if we assume that the regression between u and v is linear, and that the correlation surface of these variates is homoscedastic, we can readily obtain

$$\left.\begin{aligned} p_{21} &= r\sqrt{{}_u\beta_1}\,\sigma_u^3 \\ p_{31} &= r\,{}_u\beta_2\,\sigma_u^4 \\ p_{22} &= \{1 + r^2({}_u\beta_2 - 1)\}\,\sigma_u^4 \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(12).$$

From the first relation of (11) *bis* we can find r, since σ_w^2 and σ_u^2 have both been found, and can use this, and the known constants of the distribution of the largest variate in (12), to find the cross-product, or corrective terms of equations (11) *bis*. In this way we can find ${}_w\mu_3$ and ${}_w\mu_4$.

The legitimacy of these assumptions is, however, doubtful, because taken together they amount to an assumption of normality in the distribution of the u, v surface, which is far from correct. Mr E. S. Pearson (*Biometrika*, Vol. XVI. p. 196) has shown that equations like (12), while giving better values for p_{21}, p_{12} and p_{22} than the assumption of normality, may lead to worse results than that assumption when we substitute such approximations in equations like (11) where the higher products have *alternate* signs.

In the case of a sample of two we know the actual form of the surface, i.e. it is the part of the normal surface $z = \frac{1}{\pi} e^{-\frac{1}{2}(u^2+v^2)}$ which lies above the diagonal $u = v$, and this permits us to calculate the constants of the marginal totals which are:

$$u = -\bar{v} = {\cdot}564{,}190,$$

$$\sigma_u = \sigma_v = {\cdot}825{,}645, \quad {}_u\beta_1 = {}_v\beta_1 = {\cdot}018{,}755, \quad {}_u\beta_2 = {}_v\beta_2 = 3{\cdot}06175,$$

agreeing with the values found in Table I, p. 366. We can find the actual values of the p's; they are:

$$p_{21} = -{\cdot}077{,}079, \qquad p_{31} = +{\cdot}622{,}273,$$
$$p_{22} = +{\cdot}696{,}036, \qquad r = {\cdot}466{,}944.$$

By equations (11) we obtain:

$${}_w\mu_2 = {\cdot}726{,}758, \qquad \sigma_w = {\cdot}85250,$$
$${}_w\mu_3 = {\cdot}616{,}636, \qquad {}_w\beta_1 = {\cdot}99057,$$
$${}_w\mu_4 = 2{\cdot}043{,}625, \qquad {}_w\beta_2 = 3{\cdot}8692.$$

If we use equations (12) we have:

$p_{21} = +{\cdot}035{,}992$ instead of $-{\cdot}077{,}079$,

$p_{31} = +{\cdot}664{,}366$ „ $+{\cdot}622{,}273$,

$p_{22} = +{\cdot}673{,}602$ „ $+{\cdot}696{,}036$.

Hence from (11) *bis* we deduce

${}_w\mu_3 = -{\cdot}061{,}792$ instead of ${\cdot}616{,}636$,

${}_w\mu_4 = 1{\cdot}572{,}278$ „ $2{\cdot}043{,}625$,

${}_w\beta_1 = {\cdot}0099$ „ ${\cdot}9906$,

${}_w\beta_2 = 2{\cdot}9768$ „ $3{\cdot}8692$.

These results have been given to indicate that (12) cannot be trusted for very small samples. There is no approach to linearity. For $n = 2$ the regression curve can be plotted from the table of the probability integral. It is a hyperbolic looking curve asymptoting to $u = 0$ and $u = v$. Neither homoscedasticity nor linearity of regression is true. The correlation however of u and v rapidly diminishes and r is very small for a sample of 20. Whether the assumptions which lead to (12) give better values for β_1 and β_2 for the distribution of ranges between $n = 2$ and $n = 20$, I am not prepared to say. The hypothesis of zero correlation and the assumptions involved in (12) are shown for a series of values in Table IV, under the columns (a) and (b) respectively. Diagrams VI, VII and VIII illustrate the results of Table IV graphically, and from them the standard deviation, and β_1 and β_2, can be read off for other values of n. The nature of the early part of the β_1, β_2 curves cannot be predicted without special investigation.

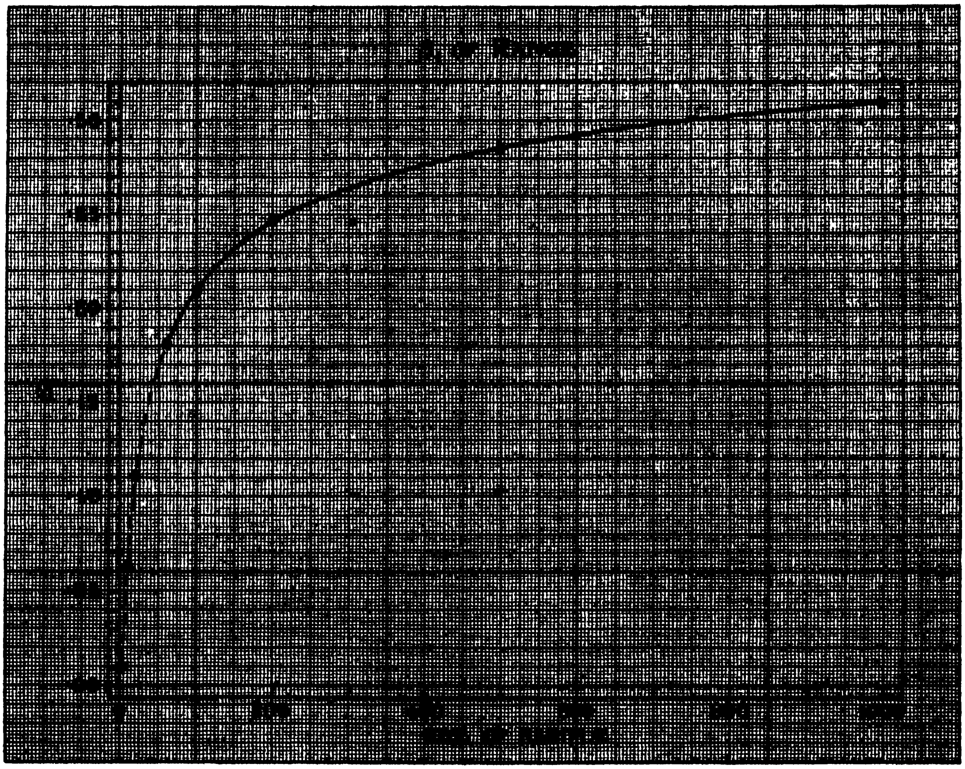


Diagram VII.

For the small samples ($n = 10$ and $n = 20$), the corrective terms are fairly considerable, and little stress is to be laid on the values given for β_1 and β_2, although it will be seen that they give fits to experimental data which are not unreasonable. For samples of 60 or more, the cross-product terms are small, and the values of β may be taken as correct for most practical purposes. It must be remembered that the third and fourth moments have large probable errors, which fact indicates that the form of distribution is not extremely sensitive to them, so that quite approximate values often give good fits.

TABLE IV.

*Constants of Distribution of Ranges**.

Size of Sample n	Standard Deviation	β_1		β_2		r
		(a)	(b)	(a)	(b)	
2	·853	[·991†]	·010	[3·869†]	2·977	·467
10	·797	·084	·063	3·166	3·138	·078
20	7·29	·125	·111	3·234	3·217	·037
60	·639	·188	·181	3·338	3·330	·012
100	·605	·215	·210	3·382	3·377	·007
200	·566	·247	·247	3·438	3·438	·00(0)
500	·524	·285	·285	3·502	3·502	·00(0)
1000	·497	·309	·309	3·544	3·544	·00(0)

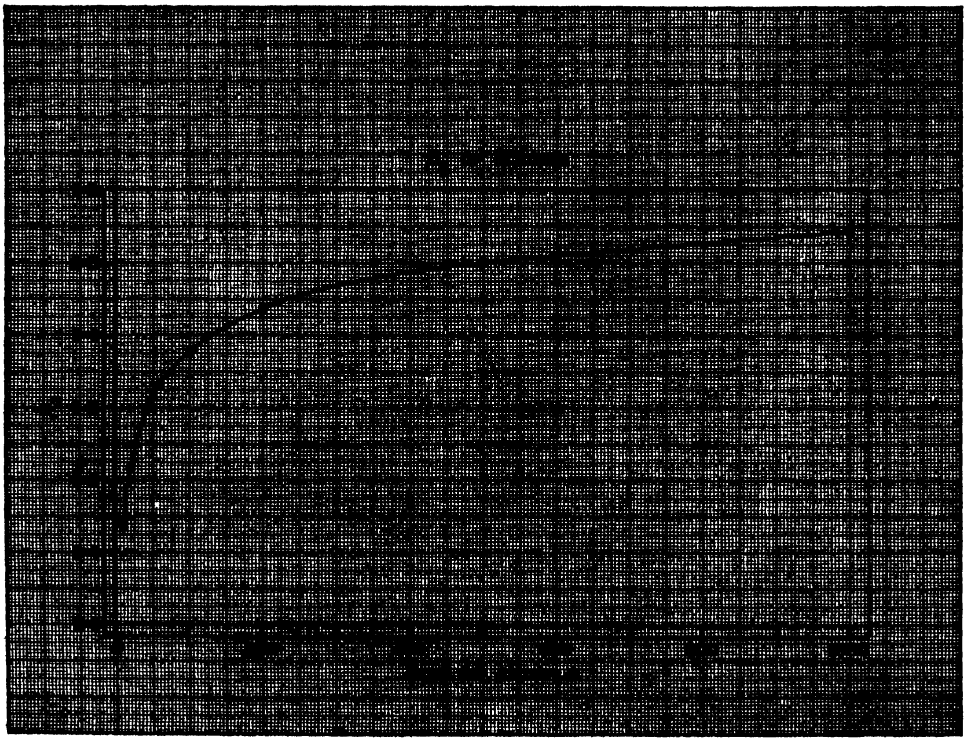


Diagram VIII.

---

* Excepting for the Standard Deviation, little reliance can be placed on the last figure of the quantities given in this table.

† Exact values.

It will be seen from Table IV that the curves to be fitted are Type VI, excepting for $n = 10$ and $n = 20$, for which a Type IV is appropriate. One would expect the true curve to be one limited in the negative direction to zero range (since negative ranges are impossible) and extended to infinity in the positive direction. These curves are, however, very near to Type V, which does satisfy these conditions, while the β's are obtained by approximative assumptions and so are not exact.

The constants for samples of less than 10 have not been attempted, except in the case $n = 2$, which has been treated by Irwin (5) and is considered more in detail here.

IV. DETERMINATION OF STANDARD DEVIATION BY THE MEAN RANGE.

Since the mean range given in the tables is in terms of the standard deviation of the original population, in any given case, the latter can be found by taking samples, determining the mean range, and dividing it by the value given by the table. This method is rapid, but subject to larger sampling errors than the moment method. It is similar to one given by Pearson (6), in which the sample is ranked, and the difference between two certain individuals measured, and divided by the value for a population having unit standard deviation. He found that best results were obtained by taking individuals near the quindeciles (those $n/15$ from each end). The standard errors of the three methods are compared in Table V. Since the distributions are skew, this quantity has no precise meaning, but it will give a good appreciation of the accuracy of the method.

The method of ranges gives better results when the samples are many and small than when they are few and large. Thus, ten samples of 10 give the standard deviation with a standard error of 8 $^{\circ}/_{\circ}$, while one sample of 100 gives it with a standard error of 12 $^{\circ}/_{\circ}$.

It is very clear that the method of ranges is markedly inferior to the method of moments, and, sample for sample, to the method of ranking. It has the advantage however that it is easy to apply, even for large samples.

TABLE V.

Standard Errors of Standard Deviation of Normal Population (Comparison of Various Methods).

Size of Sample n	Method of Moments	Method of Areas, taking Quindeciles	Method of Ranges		
			One sample of n	Five samples of n	Ten samples of n
10	·2236	·2769	·2590	·1158	·0819
20	·1581	·1958	·1952	·0873	·0617
60	·0912	·1130	·1377	·0616	·0436
100	·0707	·0876	·1207	·0540	·0382
200	·0500	·0619	·1030	·0461	·0326
500	·0316	·0391	·0862	·0386	·0273
1000	·0224	·0277	·0767	·0343	·0243

V. EXPERIMENTAL VERIFICATION.

It was considered worth while to make some sampling experiments in order to illustrate and confirm the results of this paper.

A population of 1000 individuals was made by distributing them among 27 categories according to the normal law, and with unit standard deviation. The width of each subrange was $\cdot 2\sigma$. The individuals were very small cards, which were drawn one at a time from a bag, the card being replaced and the contents of the bag well mixed between each draw. In this way, 5000 draws were made, so that one had effectively a sample of 5000 from an infinite normal population. These were compounded to give 500 samples of 10, and 500 samples of 20. These experiments gave for the mean range $2\cdot 9452 \pm \cdot 0240$ for $n = 10$ (theoretical value $3\cdot 07751$), and $3\cdot 6034 \pm \cdot 0220$ for $n = 20$ (theoretical value $3\cdot 73495$). The discrepancies between theory and practice are too great to be attributed to ordinary errors of random sampling, so it was concluded that the mixing between each draw had not been sufficient, and there was a tendency for neighbouring draws to be alike, leading to a low value of the mean range. The fact that later and more careful experiments removed the discrepancies makes this explanation plausible. The mean range thus supplies a fairly sensitive test of the randomness of a series of observations.

For the second experiment, a population of 10,000 individuals was made in a similar way, but was divided into 69 classes, the breadth of each class being $\cdot 1\sigma$. The individuals were numbered from 0000 to 9999, those in the same class having consecutive numbers; and a key was constructed, so that given a number, one could find the class to which the individual bearing that number belonged. Then 40,000 digits were taken at random from census reports, and combined by fours to give 10,000 numbers taken at random from 0000 to 9999 inclusive*. By means of the key, these numbers were converted into a random sample of the original population, which was then divided up to give 1000 samples of 5 (only half were taken), 1000 of 10 and 500 of 20.

In order to find the range, it was assumed that the individuals were concentrated at the means of the groups. Table VI gives the distributions found, the experimental groups having been combined in pairs. The theoretical distributions with which they are compared were found by fitting Type IV curves, using the constants given in columns (*b*) of Table IV. In most cases there is good agreement between the theoretical and experimental values of the constants, the most serious discrepancies being in the case of the mean for $n = 5$ (the difference is about $2\cdot 9$ times the probable error) and the standard deviation for $n = 10$ (the difference is about $2\cdot 2$ times the probable error). The theoretical curves fit the data moderately well ($n = 10$: $P = \cdot 119$, $n = 20$: $P = \cdot 455$), indicating that even for the smaller samples, the values of β_1 and β_2 given in Table IV are not unreasonable, and are sufficient for practical purposes. In Diagrams IX and X, the theoretical curve and histogram of

---

* By means of a suitable key, these numbers can be used to construct samples from any population. It is hoped that they will be published later.

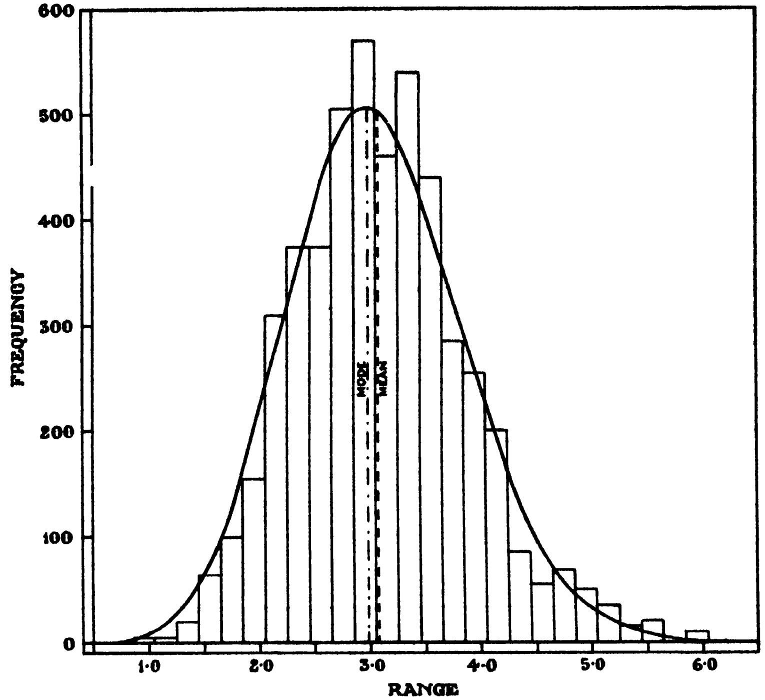

DISTRIBUTION OF RANGE IN SAMPLES OF 10

Diagram IX.

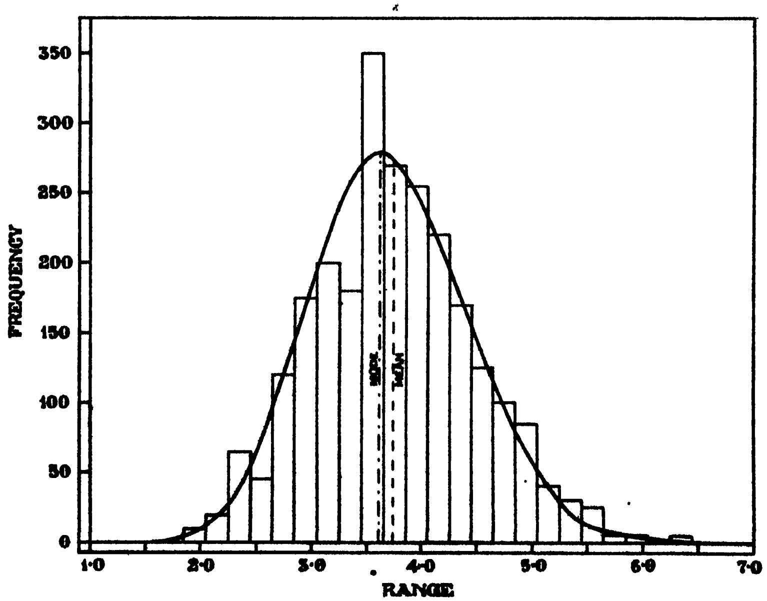

DISTRIBUTION OF RANGE IN SAMPLES OF 20

Diagram X.

the experimental data for the two cases have been superimposed. The equations to the curves are

$$n = 10: \ y = y_0 \, e^{158 \cdot 745 \, \theta} \cos^{145 \cdot 103} \theta, \quad x = 6 \cdot 3614 \tan \theta, \quad \log y_0 = \overline{30} \cdot 26265.$$

Origin at $-7·0567$ to left of mean; mode, 2·9803.

$$n = 20: \ y = y_0 \, e^{294 \cdot 086 \, \theta} \cos^{127 \cdot 100} \theta, \quad x = 3 \cdot 1734 \tan \theta, \quad \log y_0 = \overline{96} \cdot 65328.$$

Origin at $-7·4748$ to left of mean; mode, 3·6173.

TABLE VI.

Distribution of Ranges.

Range	$n=10$		$n=20$		$n=5$
	Experimental Frequency	Theoretical Frequency	Experimental Frequency	Theoretical Frequency	Experimental Frequency
·05— ·25	—	—	—	—	1
·25— ·45	—	—	—	—	4
·45— ·65	—	—	—	—	8
·65— ·85	—	—	—	—	15
·85—1·05	1 } 19	27·9	—		23
1·05—1·25	1		—	—	44
1·25—1·45	4		—	—	62
1·45—1·65	13		—	—	69
1·65—1·85	20	24·6	—	—	66
1·85—2·05	31	39·3	2 } 19	13·7	73
2·05—2·25	62	56·0	4		97
2·25—2·45	75	71·7	13		94
2·45—2·65	75	86·1	9	14·4	80
2·65—2·85	101	97·1	24	23·4	73
2·85—3·05	114	101·1	35	34·3	66
3·05—3·25	92	98·5	40	44·3	48
3·25—3·45	108	90·9	36	51·7	61
3·45—3·65	88	78·7	70	55·5	30
3·65—3·85	57	64·2	54	54·9	32
3·85—4·05	51	50·6	51	50·9	18
4·05—4·25	40	38·7	44	43·6	14
4·25—4·45	17	26·6	34	34·7	6
4·45—4·65	11	18·0	25	26·4	6
4·65—4·85	13	12·2	20	19·0	3
4·85—5·05	10	7·4	17	13·6	2
5·05—5·25	7 } 16	10·4	8	8·6	4
5·25—5·45	3		6 } 14	11·0	1
5·45—5·65	4		5		—
5·65—5·85	—		1		—
5·85—6·05	2		1		—
6·05—6·25	—	—	—		—
6·25—6·45	—	—	1		—
Totals	1000	1000·0	500	500·0	1000
Mean	3·1016 ± ·0170	3·07751	3·7508 ± ·0220	3·73495	2·3798 ± ·0189
σ	·771 ± ·012	·797	·731 ± ·0155	·729	·886 ± ·0134
β_1	·184 ± ·042	·063	·063 ± ·064	·111	·100
β_2	3·397 ± ·156	3·138	3·026 ± ·248	3·217	2·904
χ^2	25·248		14·949		—
P	·119		·455		—

Table VII gives the correlation surface between the two extreme individuals in samples of 10. A symmetrical table has been formed by counting each extreme twice, as the first and as the last member. The constants given below the table

TABLE VII.

Extremes of Sample. $n = 10$.

Central Values of First Variate (in terms of σ).

Central Values of Last Variate (in terms of σ).

	3·45	*3·25*	*3·05*	*2·85*	*2·65*	*2·45*	*2·25*	*2·05*	*1·85*	*1·65*	*1·45*	*1·25*	*1·05*	*0·85*	*0·65*	*0·45*	*0·25*	*0·05*	*−0·15*	Totals
−3·45	—	—	—	—	—	—	1	1	1	—	6	1	—	—	—	—	—	—	—	10
−3·25	—	—	—	—	1	—	1	—	2	—	—	—	—	—	—	—	—	—	—	4
−3·05	—	—	—	—	—	—	2	1	—	3	—	—	1	—	1	—	—	—	—	8
−2·85	—	—	—	—	1	1	1	3	1	4	2	3	1	3	2	1	—	1	—	24
−2·65	—	1	—	1	—	1	2	4	3	7	8	4	4	7	4	2	1	1	—	50
−2·45	—	—	—	1	1	4	4	4	4	12	7	9	8	9	8	3	1	1	—	76
−2·25	1	1	2	1	2	4	4	9	13	9	19	19	23	4	3	2	2	1	—	119
−2·05	1	—	1	3	4	4	9	14	26	24	26	28	16	14	10	6	2	—	—	188
−1·85	1	2	—	1	3	4	13	26	24	20	33	27	30	25	9	7	—	—	1	226
−1·65	—	—	3	4	7	12	9	24	20	34	33	40	36	18	13	4	1	1	—	259
−1·45	6	—	—	2	8	7	19	26	33	33	42	27	25	21	12	6	1	1	—	269
−1·25	1	—	—	3	4	9	19	28	27	40	27	20	38	20	9	1	3	1	—	250
−1·05	—	—	1	1	4	8	23	16	30	36	25	38	22	12	9	3	—	—	—	228
−0·85	—	—	—	3	7	9	4	14	25	18	21	20	12	6	3	1	—	—	—	143
−0·65	—	—	1	2	4	8	3	10	9	13	12	9	9	3	4	1	1	—	—	89
−0·45	—	—	—	1	2	3	2	6	7	4	6	1	3	1	1	—	—	—	—	37
−0·25	—		—	—	1	1	2	2	—	1	1	3	—	—	1	—	—	—	—	12
−0·05	—	—	—	1	1	1	1	—	—	1	1	1	—	—	—	—	—	—	—	7
+0·15	—	—	—	—	—	—	—	—	1		—	—	—	—	—	—	—	—	—	1
Totals	10	4	8	24	50	76	119	188	226	259	269	250	228	143	89	37	12	7	1	2000

$r = \cdot1188 \pm \cdot0210$, $\eta = \cdot141$.
(·078)

Mean $= 1\cdot5494 \pm \cdot0086$, $\sigma = \cdot5735 \pm \cdot0061$, $\beta_1 = \cdot082 \pm \cdot034$, $\beta_2 = 3\cdot079 \pm \cdot147$.
(1·53875) (·5868) (·168) (3·331)

are the coefficient of correlation r, the correlation ratio η, and the usual constants of the distribution of the first variate. The corresponding theoretical values are given in brackets. In the case of r, there is no significant discrepancy, and moreover η does not differ significantly from it, so that the non-linearity of the regression is not apparent in the sample. As for the constants of the first variate distribution (which are given by the marginal totals), there is good agreement for the mean, while for the standard deviation, and for β_1 and β_2, the difference between the theoretical and experimental values is between two and three times the probable error, a difference which is scarcely significant, but still rather disappointingly large.

In Table VIII, the results for samples of 20 are given. The correlation coefficient does not differ significantly from the theoretical value, nor from zero, while η^2 does not differ from $\bar{\eta}^2$ (the square of the correlation ratio for zero association) by as much as the probable error. Thus, the sample gives no evidence of association

TABLE VIII.

Correlation between Extremes in Samples of 20.

First Extreme.

Last Extreme.	*0·45*	*0·65*	*0·85*	*1·05*	*1·25*	*1·45*	*1·65*	*1·85*	*2·05*	*2·25*	*2·45*	*2·65*	*2·85*	*3·05*	*3·25*	*3·45*	Totals
−3·45	—	—	—	—	—	10	2	2	3	3	—	—	—	—	1	—	21
−3·25	—	—	—	—	—	—	—	4	—	3	—	2	—	—	—	1	10
−3·05	—	—	—	1	—	—	6	3	2	3	1	—	—	—	—	—	16
−2·85	—	—	1	1	3	5	10	6	10	6	4	2	—	—	—	—	48
−2·65	-	2	2	5	9	16	13	9	18	9	6	4	2	—	2	-	97
−2·45	—	3	3	9	12	17	22	20	20	14	10	6	4	1	—	-	141
−2·25	—	1	—	5	20	27	32	39	31	16	14	9	6	3	3	3	209
−2·05	1	—	8	14	21	36	38	52	48	31	20	18	10	2	—	3	302
−1·85	—	-	1	20	23	36	49	60	52	39	20	9	6	3	4	2	324
−1·65	—	3	4	17	24	36	42	49	38	32	22	13	10	6	—	2	298
−1·45		1	5	14	18	24	36	36	36	27	17	16	5	—	—	10	245
−1·25	—	—	3	9	8	18	24	23	21	20	12	9	3	—	—	—	150
−1·05	—	1	1	2	9	14	17	20	14	5	9	5	1	1	—	—	99
−0·85	—	—	—	1	3	5	4	1	8	—	3	2	1	—	-	—	28
−0·65	—	—	—	1	—	1	3	—	—	1	3	2	—	—	—	—	11
−0·45	-	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	1
Totals	1	11	28	99	150	245	298	324	302	209	141	97	48	16	10	21	2000

$r = -{\cdot}015 \pm {\cdot}021$, $\eta^2 = {\cdot}0084$, $\eta^2 = {\cdot}0070$.
(·037)

Mean $= 1{\cdot}8758 \pm {\cdot}0076$, $\sigma = {\cdot}505 \pm {\cdot}005$, $\beta_1 = {\cdot}134 \pm {\cdot}047$, $\beta_2 = 3{\cdot}245 \pm$ 177.
(1·8675) (·525) (·251) (3·469)

between the extreme individuals, as indeed one would expect; the theoretical correlation is too small to show up in such a sample. The constants of the distribution of the first variate are, on the whole, in accordance with the theoretical values*.

A further experiment was made to see if the mean range, as found for a normal population, is very different from that of a skew population. The distribution chosen was

$$y = \frac{10{,}000}{2{\cdot}8284} e^{-x} x^7, \text{ having } \sigma = 9{\cdot}988,\ \beta_1 = {\cdot}5000,\ \beta_2 = 3{\cdot}75.$$

It was divided into 76 categories, and by means of a new key, was sampled by the numbers used previously. The mean range, as given by 460 samples of 10, was 29·698 ± ·248, while the value for a normal distribution would be 30·738. The difference is about 4·2 times the probable error, and would not be significant in 100 samples. This is a very skew distribution, so that for many practical purposes one may assume the mean range for the normal curve, even when the population

---

[*] The samples of twenty were formed by combining those of ten in pairs, the first and second, the third and fourth, and so on. To give smoother regression, the other pairs (the second and third, the fourth and fifth, and so on) were also combined, so that 2000 samples of ten gave 2000 samples of twenty. These, however, are not independent, so that the probable errors (calculated for 2000) are under-estimated. Indeed, the addition of the second thousand made practically no difference to the distribution in the marginal totals; thus it may be held, that the probable errors should have been calculated for a sample of near 1000 rather than of one of 2000.

sampled departs markedly from normality. It is not suggested, however, that serious discrepancies would not occur if the results were applied to curves of U or J type, or to curves having very limited range.

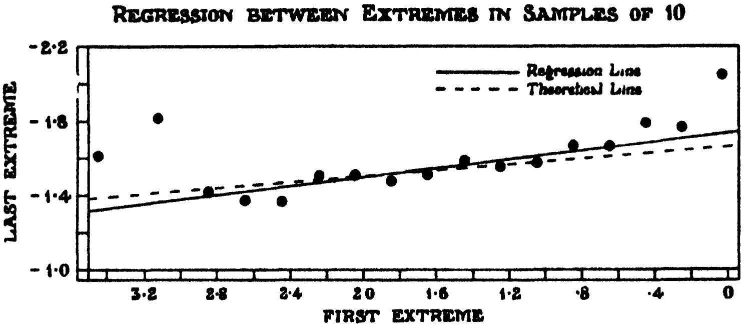


Diagram XI.

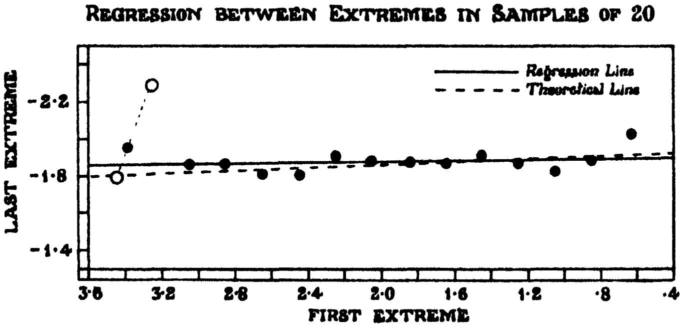


Diagram XII.

I gladly take this opportunity of thanking Professor Pearson, who has supervised the whole of this work, and has always been ready with kindly help and criticism. I am also grateful to Miss I. McLearn for preparing the diagrams for this paper, and to the British Cotton Industry Research Association for a studentship during the holding of which the present investigation was carried out.

REFERENCES.

(1) K. PEARSON. "Note on Francis Galton's Difference Problem." *Biometrika*, Vol. I. 1902, pp. 390—399.

(2) L. v. BORTKIEWICZ. "Variationabreite und mittlerer Fehler." *Sitzungsberichte der Berliner Math. Gesellschaft*, 26 October 1921.

(3) E. L. DODD. "The Greatest and Least Variate under General Laws of Error." *Trans. Amer. Math. Soc.*, Vol. XXV. October 1923, pp. 525—539.

(4) J. SPŁAWA-NEYMAN. "Sur les valeurs théoriques de la plus grande de n erreurs." *Prace Matematyczno-Fizyczne*, Vol. XXXIII.

(5) J. O. IRWIN. "The Further Theory of Francis Galton's Individual Difference Problem." *Biometrika*, Vol. XVII. 1925, pp. 100—128.

(6) K. PEARSON (Editorial). "On the Probable Errors of Frequency Constants." *Biometrika*, Vol. XIII. 1921, pp. 113—132.

TABLE IX.

Probability Integral of Distribution of Largest Individual in Samples of Size n taken from Normal Population.

x	$n=3$	$n=5$	$n=10$	x	$n=20$	$n=30$	$n=50$
−2·6	·00000 01	—	—	−0·1	·00000 01	—	—
−2·4	·00000 06	—	—	0·0	·00000 10	—	—
−2·2	·00000 27	—	—	0·1	·00000 44	·00000 00	—
−2·0	·00001 18	·00000 00	—	0·2	·00001 81	·00000 01	—
−1·8	·00004 61	·00000 01	—	0·3	·00006 58	·00000 05	—
−1·6	·00016 46	·00000 05	—	0·4	·00021 40	·00000 31	—
−1·4	·00052 67	·00000 34	—	0·5	·00062 43	·00001 56	·00000 00
−1·2	·00152 36	·00002 02	—	0·6	·00164 32	·00006 66	·00000 01
−1·0	·00399 36	·00010 05	·00000 00	0·7	·00392 45	·00024 59	·00000 10
−0·8	·00950 86	·00042 68	·00000 02	0·8	·00855 30	·00079 10	·00000 68
−0·6	·02062 79	·00155 15	·00000 24	0·9	·01710 66	·00223 74	·00003 83
−0·4	·04091 32	·00485 78	·00002 36	1·0	·03158 49	·00561 33	·00017 73
−0·2	·07448 05	·01318 47	·00017 38	1·1	·05415 41	·01260 22	·00068 25
0·0	·12500 00	·03125 00	·00097 66	1·2	·08673 24	·02554 30	·00221 54
+0·2	·19436 59	·06521 79	·00425 34	1·3	·13051 91	·04715 32	·00615 44
+0·4	·28155 45	·12094 95	·01462 88	1·4	·18561 36	·07996 77	·01484 31
+0·6	·38225 71	·20133 81	·04053 70	1·5	·25085 78	·12564 39	·03151 87
+0·8	·48957 33	·30110 92	·09248 24	1·6	·32395 21	·18438 31	·05973 13
+1·0	·59555 51	·42157 02	·17772 15	1·7	·40180 81	·25470 01	·10234 03
+1·2	·69299 04	·54268 20	·29450 37	1·8	·48102 59	·33362 05	·16048 01
+1·4	·77676 83	·65637 57	·43082 90	1·9	·55836 69	·41723 38	·23296 96
+1·6	·84444 64	·75443 22	·56916 79	2·0	·63112 07	·50138 19	·31643 25
+1·8	·89603 56	·83280 27	·69356 03	2·1	·69731 48	·58229 54	·40604 32
+2·0	·93329 05	·89130 86	·79443 10	2·2	·75576 85	·65702 70	·49656 03
+2·2	·95886 69	·93238 91	·86934 95	2·3	·80602 50	·72364 03	·58327 22
+2·4	·97560 85	·95967 88	·92098 35	2·4	·84821 05	·78118 79	·66261 18
+2·6	·98608 15	·97691 03	·95435 38	2·5	·88286 71	·82955 08	·73238 31
+2·8	·99235 42	·98728 95	·97474 05	2·6	·91079 11	·86921 69	·79167 51
+3·0	·99595 58	·99326 87	·98658 27	2·7	·93289 10	·90105 41	·84059 12
+3·2	·99794 00	·99656 90	·99314 98	2·8	·95011 90	·92611 95	·87992 38
+3·4	·99898 95	·99831 65	·99663 58	2·9	·96333 78	·94551 29	·91084 93
+3·6	·99952 27	·99920 47	·99841 01	3·0	·97334 55	·96028 58	·93468 99
+3·8	·99978 30	·99963 83	·99927 65	3·1	·98082 48	·97137 55	·95274 92
+4·0	·99990 50	·99984 17	·99968 33	3·2	·98634 66	·97958 99	·96621 52
+4·2	·99996 00	·99993 10	·99986 66	3·3	99037 58	·98559 85	·97611 29
+4·4	·99998 38	·99997 29	·99994 59	3·4	·99328 29	·98994 13	·98329 19
+4·6	·99999 37	·99998 94	·99997 89	3·5	·99535 77	·99304 46	·98843 46
+4·8	·99999 76	·99999 60	·99999 21	3·6	·99682 26	·99523 77	·99207 55
+5·0	·99999 91	·99999 86	·99999 71	3·7	·99784 62	·99677 11	·99462 42
+5·2	·99999 96	·99999 95	·99999 90	3·8	·99855 40	·99783 18	·99638 90
+5·4	·99999 99	·99999 98	·99999 97	3·9	·99903 85	·99855 81	·99759 80
+5·6	1·00000 00	·99999 99	·99999 99	4·0	·99936 68	·99905 03	·99841 77
+5·8	—	1·00000 00	1·00000 00	4·1	·99958 69	·99938 05	·99896 76
				4·2	·99973 31	·99959 97	·99933 29
				4·3	·99982 92	·99974 38	·99957 31
				4·4	·99989 18	·99983 76	·99972 94
				4·5	·99993 21	·99989 81	·99983 01
				4·6	·99995 78	·99993 66	·99989 44
				4·7	·99997 40	·99996 10	·99993 50
				4·8	·99998 41	·99997 62	·99996 03
				4·9	·99999 04	·99998 56	·99997 60
				5·0	·99999 43	·99999 14	·99998 57
				5·1	·99999 66	·99999 49	·99999 15
				5·2	·99999 80	·99999 70	·99999 50
				5·3	·99999 89	·99999 83	·99999 71
				5·4	·99999 93	·99999 90	·99999 83
				5·5	·99999 96	·99999 94	·99999 91
				5·6	·99999 98	·99999 97	·99999 95
				5·7	·99999 99	·99999 98	·99999 97
				5·8	·99999 99	·99999 99	·99999 98
				5·9	1·00000 00	·99999 99	·99999 99
				6·0	—	1·00000 00	1·00000 00

TABLE IX.—(*continued*).

Probability Integral of Distribution of Largest Individual in Samples of Size n taken from Normal Population.

x	$n=100$	$n=200$	$n=300$	$n=400$	$n=500$	$n=600$	$n=700$	$n=800$	$n=900$	$n=1000$
1·0	·00000 00	—	—	—	—	—	—	—	—	—
1·1	·00000 05	—	—	—	—	—	—	—	—	—
1·2	·00000 49	—	—	—	—	—	—	—		—
1·3	·00003 79	—	—	—	—	—	—	—	—	—
1·4	·00022 03	·00000 00	—	—	—	—	—	—	—	—
1·5	·00099 34	·00000 10	—	—	—	—	—	—	—	—
1·6	·00356 78	·00001 27	·00000 00	—	—	—	—	—	—	—
1·7	·01047 35	·00010 97	·00000 11	·00000 00	—	—	—	—	—	—
1·8	·02575 39	·00066 33	·00001 71	·00000 04	·00000 00	—	—	—	—	—
1·9	·05427 48	·00294 58	·00015 99	·00000 87	·00000 05	·00000 00	·00000 00	.		—
2·0	·10012 95	·01002 59	·00100 39	·00010 05	·00001 01	·00000 10	·00000 01	·00000 00	·00000 00	—
2·1	·16487 11	·02718 25	·00448 16	·00073 88	·00012 18	·00002 01	·00000 33	·00000 05	·00000 01	·00000 00
2·2	·24657 22	·06079 78	·01499 11	·00369 64	·00091 14	·00022 47	·00005 54	·00001 37	·00000 34	·00000 08
2·3	·34020 64	·11574 04	03937 56	·01339 58	·00455 74	·00155 04	·00052 75	·00017 94	·00006 10	·00002 08
2·4	·43905 44	·19276 87	·08463 60	·03715 98	·01631 52	·00716 32	·00314 51	·00138 08	·00060 63	·00026 61
2·5	·53638 50	·28770 89	·15432 27	·08277 66	·04440 01	·02381 56	·01277 43	·00685 20	·00367 53	·00197 14
2·6	·62674 94	·39281 48	·24619 64	·15430 35	·09670 96	·06061 27	·03798 90	·02380 96	·01492 26	·00935 27
2·7	·70659 35	·49927 43	·35278 40	·24927 49	·17613 60	·12445 66	·08794 02	·06213 81	·04390 63	·03102 39
2·8	·77426 58	·59948 76	·46416 27	·35938 53	·27825 98	·21544 71	·16681 33	·12915 78	·10000 25	·07742 85
2·9	·82964 65	·68831 33	·57105 67	47377 51	·39306 59	·32610 57	·27055 25	·22446 29	·18622 48	·15450 08
3·0	·87364 51	·76325 58	·66681 48	·58255 95	·50895 03	·44464 19	·38845 93	·33937 55	·29649 38	·25903 00
3·1	·90773 10	·82397 56	·74794 82	·67893 58	·61629 11	·55942 66	·50780 89	·46095 38	·41842 21	·37981 48
3·2	·93357 18	·87155 63	·81366 04	·75961 04	·70915 08	·66204 32	·61806 49	·57700 79	·53867 83	·50289 49
3·3	·95279 63	·90782 09	·86496 84	·82413 87	·78523 64	·74817 03	·71285 40	·67920 46	·64714 37	·61659 60
3·4	·96686 29	·93482 38	·90384 64	·87389 55	·84493 72	·81693 84	·78986 74	·76369 34	·73838 68	·71391 22
3·5	·97700 29	·95453 48	·93258 33	·91113 66	·89018 31	·86971.15	·84971 07	·83016 99	·81107 84	·79242 60
3·6	·98421 38	·96867 68	·95338 51	·93833 48	·92352 20	·90894 31	·89459 44	·88047 22	·86657 28	·85289 29
3·7	·98927 73	·97866 44	·96817 57	·95779 43	·94752 42	93736 43	·92731 33	·91737 00	·90753 33	·89780 22
3·8	·99279 10	·98563 41	·97852 87	·97147 45	·96447 12	·95751 83	·95061 57	·94376 27	·93695 92	·93020 46
3·9	·99520 18	·99042 66	·98567 44	·98094 49	·97623 81	97155 39	·96689 22	·96225 29	·95763 58	·95304 09
4·0	·99683 78	·99368 57	·99054 35	98741 12	·98428 88	·98117 64	·97807 37	·97498 09	·97189 79	·96882 45
4·1	·99793 64	·99587 70	99382 18	·99177 10	·98972 43	·98768 19	·98564 36	·98360 96	·98157 98	·97955 42
4·2	·99866 63	·99733 44	·99600 43	·99467 59	·99334 93	·99202 45	·99070 14	·98938 01	·98806 06	·98674 28
4·3	·99914 64	·99829 35	·99744 13	·99658 99	·99573 91	·99488 91	·99403 99	·99319 13	·99234 35	·99149 64
4·4	·99945 89	·99891 81	·99837 75	·99783 73	·99729 74	·99675 78	·99621 84	·99567 93	·99514 05	·99460 21
4·5	·99966 03	·99932 07	99898 12	·99864 19	·99830 26	·99796 35	·99762 45	·99728 56	·99694 68	·99660 81
4·6	·99978 88	·99957 76	·99936 65	·99915 54	·99894 43	·99873 33	·99852 24	·99831 15	·99810 06	·99788 98
4·7	·99986 99	·99973 99	·99960 98	·99947 98	·99934 97	·99921 98	·99908 98	·99895 99	·99883 00	·99870 00
4·8	·99992 07	·99984 13	·99976 20	·99968 27	·99960 34	·99952 41	·99944 48	·99936 55	·99928 62	·99920 70
4·9	·99995 21	·99990 42	·99985 63	·99980 83	·99976 04	·99971 25	·99966 46	·99961 67	·99956 88	·99952 09
5·0	·99997 13	·99994 27	·99991 40	·99988 53	·99985 67	·99982 80	·99979 94	·99977 07	·99974 21	·99971 34
5·1	·99998 30	·99996 60	·99994 91	·99993 21	·99991 51	·99989 81	·99988 11	·99986 42	·99984 72	·99983 02
5·2	·99999 00	·99998 01	·99997 01	·99996 01	·99995 02	·99994 02	·99993 03	·99992 03	·99991 03	·99990 04
5·3	·99999 42	·99998 84	·99998 26	·99997 67	·99997 11	·99996 53	·99995 95	·99995 37	·99994 79	·99994 21
5·4	·99999 67	·99999 33	·99999 00	·99998 67	·99998 33	·99998 00	·99997 67	·99997 33	·99997 00	·99996 67
5·5	·99999 81	·99999 62	·99999 43	·99999 24	·99999 05	·99998 86	·99998 68	·99998 48	·99998 29	·99998 10
5·6	·99999 89	·99999 79	·99999 68	·99999 57	·99999 46	·99999 36	·99999 25	·99999 14	·99999 03	·99998 93
5·7	·99999 94	·99999 88	·99999 82	·99999 76	·99999 70	·99999 64	·99999 64	·99999 52	·99999 46	·99999 40
5·8	·99999 97	·99999 93	·99999 90	·99999 87	·99999 83	·99999 80	·99999 77	·99999 74	·99999 70	·99999 67
5·9	·99999 98	·99999 96	·99999 95	·99999 93	·99999 91	·99999 89	·99999 87	·99999 85	·99999 84	·99999 82
6·0	·99999 99	·99999 98	·99999 97	·99999 96	·99999 96	·99999 94	·99999 93	·99999 92	·99999 91	·99999 90
6·1	·99999 99	·99999 99	·99999 98	·99999 98	·99999 98	·99999 97	·99999 96	·99999 96	·99999 95	·99999 95
6·2	1·00000 00	1·00000 00	·99999 99	·99999 99	·99999 99	·99999 98	·99999 98	·99999 98	·99999 98	·99999 98
6·3	—	—	1·00000 00	·99999 99	·99999 99	·99999 99	·99999 99	·99999 99	·99999 99	·99999 99
6·4	—	—	—	1·00000 00	1·00000 00	1·00000 00	1·00000 00	·99999 99	·99999 99	·99999 99
6·5	—	—	—	—	—	—	—	1·00000 00	1·00000 00	1·00000 00

TABLE X.

Mean Range of Samples of Size n taken from Normal Population
(given in terms of Standard Deviation).

n	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*n*
0	–	—	1·12838	1·69257	2·05875$^+$	2·32593	2·53441	2·70436	2·84720	2·97003	*0*
10	3·07751	3·17287	3·25846	3·33598	3·40676	3·47183	3·53198	3·58788	3·64006	3·68896	*10*
20	3·73495$^+$	3·77834	3·81938	3·85832	3·89535$^-$	3·93063	3·96432	3·99654	4·02741	4·05704	*20*
30	4·08552	4·11293	4·13934	4·16482	4·18943	4·21322	4·23625$^-$	4·25855$^+$	4·28018	4·30117	*30*
40	4·32156	4·34136	4·36063	4·37938	4·39764	4·41544	4·43279	4·44972	4·46624	4·48238	*40*
50	4·49815$^-$	4·51356	4·52864	4·54339	4·55783	4·57197	4·58582	4·59939	4·61270	4·62575$^+$	*50*
60	4·63856	4·65112	4·66346	4·67557	4·68747	4·69916	4·71065$^-$	4·72194	4·73305$^-$	4·74397	*60*
70	4·75472	4·76530	4·77570	4·78595$^+$	4·79604	4·80598	4·81578	4·82543	4·83494	4·84431	*70*
80	4·85355$^-$	4·86266	4·87165$^-$	4·88051	4·88926	4·89789	4·90641	4·91481	4·92312	4·93131	*80*
90	4·93940	4·94739	4·95529	4·96309	4·97079	4·97841	4·98593	4·99337	5·00073	5·00800	*90*
100	5·01519	5·02230	5·02933	5·03628	5·04316	5·04997	5·05670	5·06337	5·06996	5·07649	*100*
110	5·08295$^-$	5·08934	5·09568	5·10195$^-$	5·10815$^+$	5·11430	5·12039	5·12642	5·13239	5·13831	*110*
120	5·14417	5·14998	5·15573	5·16144	5·16709	5·17269	5·17824	5·18374	5·18919	5·19460	*120*
130	5·19996	5·20528	5·21055$^-$	5·21578	5·22096	5·22610	5·23120	5·23625$^+$	5·24127	5·24624	*130*
140	5·25118	5·25608	5·26094	5·26576	5·27054	5·27529	5·28000	5·28468	5·28932	5·29392	*140*
150	5·29849	5·30303	5·30754	5·31201	5·31645$^-$	5·32086	5·32523	5·32958	5·33389	5·33818	*150*
160	5·34244	5·34666	5·35086	5·35503	5·35917	5·36328	5·36737	5·37142	5·37545$^+$	5·37946	*160*
170	5·38344	5·38739	5·39132	5·39522	5·39910	5·40295$^+$	5·40678	5·41059	5·41437	5·41812	*170*
180	5·42186	5·42557	5·42926	5·43293	5·43657	5·44019	5·44380	5·44738	5·45093	5·45447	*180*
190	5·45799	5·46149	5·46497	5·46842	5·47186	5·47528	5·47868	5·48206	5·48542	5·48876	*190*
200	5·49209	5·49539	5·49868	5·50195$^-$	5·50520	5·50843	5·51165$^+$	5·51485$^+$	5·51803	5·52120	*200*
210	5·52435$^-$	5·52748	5·53060	5·53370	5·53678	5·53985$^+$	5·54291	5·54594	5·54897	5·55197	*210*
220	5·55497	5·55794	5·56091	5·56385$^+$	5·56679	5·56971	5·57261	5·57550	5·57838	5·58124	*220*
230	5·58409	5·58692	5·58975$^-$	5·59255$^+$	5·59535$^-$	5·59813	5·60090	5·60366	5·60640	5·60913	*230*
240	5·61185$^-$	5·61456	5·61725$^-$	5·61993	5·62260	5·62526	5·62790	5·63054	5·63316	5·63577	*240*
250	5·63837	5·64096	5·64353	5·64610	5·64865$^+$	5·65119	5·65373	5·65625$^-$	5·65876	5·66126	*250*
260	5·66375$^-$	5·66623	5·66869	5·67115$^+$	5·67360	5·67604	5·67847	5·68088	5·68329	5·68569	*260*
270	5·68808	5·69046	5·69282	5·69518	5·69753	5·69987	5·70221	5·70453	5·70684	5·70914	*270*
280	5·71144	5·71372	5·71600	5·71827	5·72053	5·72278	5·72502	5·72725$^+$	5·72948	5·73170	*280*
290	5·73390	5·73610	5·73829	5·74048	5·74265$^+$	5·74482	5·74698	5·74913	5·75127	5·75341	*290*
300	5·75553	5·75765$^+$	5·75977	5·76187	5·76397	5·76605$^+$	5·76814	5·77021	5·77228	5·77434	*300*
310	5·77639	5·77843	5·78047	5·78250	5·78453	5·78654	5·78855$^+$	5:79055$^+$	5·79255$^+$	5·79454	*310*
320	5·79652	5·79850	5·80046	5·80243	5·80438	5·80633	5·80827	5·81021	5·81214	5·81406	*320*
330	5·81598	5·81789	5·81979	5·82169	5·82358	5·82546	5·82734	5·82922	5·83108	5·83294	*330*
340	5·83480	5·83665$^-$	5·83849	5·84033	5·84216	5·84398	5·84580	5·84762	5·84942	5·85123	*340*
350	5·85302	5·85482	5·85660	5·85838	5·86016	5·86192	5·86369	5·86545$^-$	5·86720	5·86895$^-$	*350*
360	5·87069	5·87243	5·87416	5·87588	5·87761	5·87932	5·88103	5·88274	5·88444	5·88614	*360*
370	5·88783	5·88951	5·89119	5·89287	5·89454	5·89621	5·89787	5·89952	5·90118	5·90282	*370*
380	5·90447	5·90610	5·90774	5·90936	5·91099	5·91261	5·91422	5·91583	5·91744	5·91904	*380*
390	5·92063	5·92223	5·92381	5·92540	5·92697	5·92855$^-$	5·93012	5·93168	5·93325$^-$	5·93480	*390*
400	5·93636	5·93790	5·93945$^-$	5·94099	5·94253	5·94406	5·94558	5·94711	5·94863	5·95014	*400*
410	5·95166	5·95316	5·95467	5·95617	5·95766	5·95915$^+$	5·96064	5·96212	5·96360	5·96508	*410*
420	5·96655$^+$	5·96802	5·96949	5·97095$^-$	5·97240	5·97386	5·97531	5·97675$^+$	5·97820	5·97963	*420*
430	5·98107	5·98250	5·98393	5·98535$^+$	5·98677	5·98819	5·98960	5·99101	5·99242	5·99382	*430*
440	5·99522	5·99662	5·99801	5·99940	6·00079	6·00217	6·00355$^-$	6·00492	6·00630	6·00766	*440*
450	6·00903	6·01039	6·01175$^+$	6·01311	6·01446	6·01581	6·01716	6·01850	6·01984	6·02117	*450*
460	6·02251	6·02384	6·02516	6·02649	6·02781	6·02913	6·03044	6·03175$^+$	6·03306	6·03437	*460*
470	6·03567	6·03697	6·03826	6·03956	6·04085$^-$	6·04214	6·04342	6·04470	6·04598	6·04726	*470*
480	6·04853	6·04980	6·05107	6·05233	6·05359	6·05485$^+$	6·05611	6·05736	6·05861	6·05986	*480*
490	6·06110	6·06234	6·06358	6·06482	6·06605$^+$	6·06728	6·06851	6·06974	6·07096	6·07218	*490*

TABLE X.—(*continued*).

Mean Range of Samples of Size n taken from Normal Population (given in terms of Standard Deviation).

n	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*n*
500	6·07340	6·07461	6·07583	6·07703	6·07824	6·07945−	6·08065−	6·08185−	6·08304	6·08424	*500*
510	6·08543	6·08662	6·08781	6·08899	6·09017	6·09135+	6·09253	6·09370	6·09487	6·09604	*510*
520	6·09721	6·09837	6·09953	6·10069	6·10185+	6·10301	6·10416	6·10531	6·10646	6·10760	*520*
530	6·10874	6·10988	6·11102	6·11216	6·11329	6·11442	6·11555+	6·11668	6·11780	6·11892	*530*
540	6·12004	6·12116	6·12228	6·12339	6·12450	6·12561	6·12672	6·12782	6·12892	6·13002	*540*
550	6·13112	6·13222	6·13331	6·13440	6·13549	6·13658	6·13766	6·13874	6·13982	6·14090	*550*
560	6·14198	6·14305+	6·14413	6·14520	6·14626	6·14733	6·14839	6·14946	6·15052	6·15157	*560*
570	6·15263	6·15368	6·15474	6·15579	6·15683	6·15788	6·15892	6·15996	6·16101	6·16204	*570*
580	6·16308	6·16411	6·16515−	6·16618	6·16720	6·16823	6·16926	6·17028	6·17130	6·17232	*580*
590	6·17333	6·17435+	6·17536	6·17638	6·17738	6·17839	6·17940	6·18040	6·18140	6·18241	*590*
600	6·18340	6·18440	6·18540	6·18639	6·18738	6·18837	6·18936	6·19034	6·19133	6·19231	*600*
610	6·19329	6·19427	6·19525−	6·19623	6·19720	6·19817	6·19914	6·20011	6·20108	6·20204	*610*
620	6·20301	6·20397	6·20493	6·20589	6·20685−	6·20780	6·20876	6·20971	6·21066	6·21161	*620*
630	6·21255+	6·21350	6·21444	6·21539	6·21633	6·21727	6·21820	6·21914	6·22007	6·22101	*630*
640	6·22194	6·22287	6·22379	6·22472	6·22565−	6·22657	6·22749	6·22841	6·22933	6·23025−	*640*
650	6·23116	6·23208	6·23299	6·23390	6·23481	6·23572	6·23662	6·23753	6·23843	6·23934	*650*
660	6·24024	6·24113	6·24203	6·24293	6·24382	6·24472	6·24561	6·24650	6·24739	6·24827	*660*
670	6·24916	6·25005−	6·25093	6·25181	6·25269	6·25357	6·25445−	6·25532	6·25620	6·25707	*670*
680	6·25794	6·25881	6·25968	6·26055+	6·26142	6·26228	6·26315−	6·26401	6·26487	6·26573	*680*
690	6·26659	6·26744	6·26830	6·26915+	6·27001	6·27086	6·27171	6·27256	6·27340	6·27425+	*690*
700	6·27510	6·27594	6·27678	6·27762	6·27846	6·27930	6·28014	6·28097	6·28181	6·28264	*700*
710	6·28347	6·28430	6·28513	6·28596	6·28679	6·28762	6·28844	6·28926	6·29009	6·29091	*710*
720	6·29173	6·29254	6·29336	6·29418	6·29499	6·29581	6·29662	6·29743	6·29824	6·29905−	*720*
730	6·29985+	6·30066	6·30147	6·30227	6·30307	6·30387	6·30467	6·30547	6·30627	6·30707	*730*
740	6·30786	6·30866	6·30945+	6·31024	6·31103	6·31182	6·31261	6·31340	6·31419	6·31497	*740*
750	6·31576	6·31654	6·31732	6·31810	6·31888	6·31966	6·32044	6·32121	6·32199	6·32276	*750*
760	6·32353	6·32431	6·32508	6·32585−	6·32661	6·32738	6·32815−	6·32891	6·32968	6·33044	*760*
770	6·33120	6·33197	6·33273	6·33348	6·33424	6·33500	6·33575+	6·33651	6·33726	6·33802	*770*
780	6·33877	6·33952	6·34027	6·34102	6·34176	6·34251	6·34325+	6·34400	6·34474	6·34548	*780*
790	6·34623	6·34697	6·34771	6·34844	6·34918	6·34992	6·35065+	6·35139	6·35212	6·35285+	*790*
800	6·35358	6·35431	6·35504	6·35577	6·35650	6·35722	6·35795−	6·35867	6·35940	6·36012	*800*
810	6·36084	6·36156	6·36228	6·36300	6·36372	6·36433	6·36515+	6·36587	6·36658	6·36729	*810*
820	6·36800	6·36872	6·36942	6·37013	6·37084	6·37155−	6·37226	6·37296	6·37367	6·37437	*820*
830	6·37507	6·37577	6·37647	6·37717	6·37787	6·37857	6·37927	6·37997	6·38066	6·38136	*830*
840	6·38205−	6·38274	6·38343	6·38412	6·38482	6·38550	6·38619	6·38688	6·38757	6·38825+	*840*
850	6·38894	6·38962	6·39031	6·39099	6·39167	6·39235−	6·39303	6·39371	6·39438	6·39506	*850*
860	6·39574	6·39641	6·39709	6·39776	6·39843	6·39911	6·39978	6·40045−	6·40112	6·40179	*860*
870	6·40245+	6·40312	6·40379	6·40445+	6·40512	6·40578	6·40644	6·40711	6·40777	6·40843	*870*
880	6·40909	6·40975−	6·41040	6·41106	6·41172	6·41237	6·41303	6·41368	6·41433	6·41499	*880*
890	6·41564	6·41629	6·41694	6·41759	6·41824	6·41889	6·41953	6·42018	6·42082	6·42147	*890*
900	6·42211	6·42276	6·42340	6·42404	6·42468	6·42532	6·42596	6·42660	6·42723	6·42787	*900*
910	6·42851	6·42914	6·42978	6·43041	6·43104	6·43168	6·43231	6·43294	6·43357	6·43420	*910*
920	6·43483	6·43546	6·43608	6·43671	6·43734	6·43796	6·43858	6·43921	6·43983	6·44045+	*920*
930	6·44108	6·44170	6·44232	6·44294	6·44355+	6·44417	6·44479	6·44540	6·44602	6·44664	*930*
940	6·44725+	6·44786	6·44848	6·44909	6·44970	6·45031	6·45092	6·45153	6·45214	6·45275−	*940*
950	6·45335+	6·45396	6·45457	6·45517	6·45578	6·45638	6·45698	6·45759	6·45819	6·45879	*950*
960	6·45939	6·45999	6·46059	6·46119	6·46178	6·46238	6·46298	6·46357	6·46417	6·46476	*960*
970	6·46536	6·46595−	6·46654	6·46714	6·46773	6·46832	6·46891	6·46950	6·47008	6·47067	*970*
980	6·47126	6·47185−	6·47243	6·47302	6·47360	6·47419	6·47477	6·47535+	6·47594	6·47652	*980*
990	6·47710	6·47768	6·47826	6·47884	6·47942	6·47999	6·48057	6·48115−	6·48172	6·48230	*990*
1000	6·48287	—	—	—	—	—	—	—	—	—	*1000*

BAYES' THEOREM, EXAMINED IN THE LIGHT OF EXPERIMENTAL SAMPLING.

BY EGON S. PEARSON, M.A.

1. IN reading from the long succession of papers, and chapters in books on the theory of probability that have been written on the problem of "inverse probabilities" during the last century and a half, it is difficult not to be struck by the absence of any systematic attempt to put the theoretical rules to the test. Both the supporters and detractors of what has been termed Bayes' Theorem have relied almost entirely on the logic of their argument; this has been so from the time when Price*, communicating Bayes' notes to the Royal Society, first dwelt on the definite rule by which a man fresh to this world ought to regulate his expectation of succeeding sunrises, up to recent days when Keynes† has argued that it is almost discreditable to base any reliance on so foolish a theorem. Such appeals to experience as have been made are of little final value; Venn‡ it is true realised that the rule could not be tested by single instances but only in the long run of experience, yet after giving three unfavourable cases he was content to leave matters with the assurance that the experimenter "could hardly fail" to find himself "grossly wrong" if he based his expectation on the rule in further trials. The main reason why the critics have failed to produce any solid evidence in favour of their contentions, is perhaps because they have felt that the "equal distribution of ignorance" and the "distribution of *a priori* probabilities" are conceptions too intangible and unreal to be brought to the test of numbers, but the supporters of the theory can have no excuse for leaving matters as they stand except the plea of the time and labour involved in collecting data. If Bayes' Theorem is of practical value to the statistician it must be possible to show by a series of *ad hoc* experiments that the predictions which it makes are in fact justified. My excuse therefore in returning to this much and long discussed subject is to describe the results of a long series of experimental samplings which have been carried out at intervals during the last four years with a view to making more clear the use and the limitations of Bayes' Theorem in the field of practical statistics. But before dealing with the experiments it will be helpful to consider certain aspects of the problem from the theoretical point of view.

According to Price§, Bayes originally reached his result by a " very ingenious solution" based on the supposition that " the chance was the same that the probability for the happening of an event perfectly unknown should lie between any two equidistant degrees." But afterwards Bayes "considered that the postulate on which he had argued might not perhaps be looked upon by all as reasonable; and therefore he chose to lay down in another form the proposition in which he thought the solution of the problem is contained." That is to say after first looking at the problem from the point of view of the distribution of chances, he developed the proof which made use of the idea of balls dropped on a table. A first ball is dropped which is equally likely to rest anywhere on the table, and this is followed by the dropping of n other balls also equally likely to rest anywhere, of which p fall on one side of the first ball and q on the other. Contained in this proof is the idea of an event depending on the value of some underlying variate; happening when the variate is above a limiting value, failing to happen when it is below this value. This method of approach was passed over by the later writers, who with Laplace started simply from the idea of an equal distribution of chances, and until recently the controversy that has raged has been concerned almost entirely with what Venn termed the Rule of Succession*. This rule concerns itself with the chance of an event reappearing in a *single* trial only; it is usually given as follows. If an event has been observed to occur p times out of n trials under certain conditions, then the probability of its recurrence under the same conditions is $(p+1)/(n+2)$. The rule was sometimes given with $p=0$, and applied occasionally solely to the prediction of events which had never been known to occur before; it is partly this one-sided application of the theorem which has been cited as bringing discredit on the whole theory of inverse probabilities.

---

* *Phil. Trans.* Vol. LIII. 1763, p. 370 *et seq.*

† *A Treatise on Probability*, 1921, p. 382.

‡ *The Logic of Chance*, 1876, p. 180.

§ *loc. cit.*

Bayes' conception that the happening of an event depends on the value of an underlying variate has recently been extended by K. Pearson in this Journal†. The assumptions involved in his method of approach are put out clearly in the last of these papers but the position may be summarised as follows. A certain character observed among the individuals of a population can be considered as dependent on some underlying variate, x, such that the character appears when x exceeds a limiting value, ξ, but fails to appear when x is less than ξ. We are then concerned with two functions:

$f(x)$, the frequency curve for values of x in the population,

$\phi(\xi)$, the frequency curve giving the *a priori* possible values of ξ.

To reach the final result it is necessary to assume that the functions f and ϕ are identical. This would be the case naturally in a simple extension of Bayes' idea of balls on a billiard table, for here the factors determining the position of the first ball dropped would be the same as those determining that of each of the subsequent balls‡, but in general this correspondence does not exist and it seems to me often difficult to feel confident that it is justifiable to link f with ϕ.

---

* *loc. cit.* p. 176.

† *Biometrika*, Vol. XIII. pp. 1—16; 300, 301. Vol. XVI. pp. 190—193.

‡ That is to say, if we may suppose that the presence of previously dropped balls does not affect the rolling of those that follow.

A concrete example will perhaps illustrate the difficulty. Suppose we are considering a population of human beings in which the character under consideration is "a temperature over 100° F." This temperature of the body depends in some way on its functioning, on some factor of metabolism; call this factor x and suppose that its distribution among the individuals is represented by $f(x)$. At a certain value ξ of the metabolic factor, the temperature in the individual rises to 100° F.; we do not know this value ξ nor the distribution of possible values but we suppose it to be $\phi(\xi)$. Then in order to reach Bayes' Theorem we assume that $f(x) = c.\phi(x)$. Can we feel confident that this is a reasonable assumption? The frequency distribution of x is something perfectly definite, if unknown to us, depending entirely on the factors influencing temperature. But the value of ξ depends on the arbitrary temperature limit that we have fixed; we might have chosen 99° F. or 150° F. and should have had the same justification in applying Bayes' Theorem in each case. Can we then relate ϕ to f, and are we really justified in assuming as a general rule that the "most likely" value of ξ is the modal value of x (which follows if $f = c.\phi$)? We cannot appeal directly to experience, for no method of practical sampling would I think enable us to determine the form and relationship of f and ϕ. But this should be noted; if $F(P)$ is a function representing the distribution of probabilities or proportions in the world of experience, it has been shown* that the relation $f = c.\phi$ necessarily implies $F =$ constant. Now this distribution $F(P)$ is of a more tangible nature and can be made the subject for experimental inquiry. I prefer therefore to approach the problem from a rather different point of view, an approach which appears to lead up more naturally to the series of experimental samplings to be discussed, than one which involves the idea of the f and ϕ functions.

2. Suppose a man to be surrounded with a number, L, of bags each containing a large number of black and white balls. In a particular bag the proportion of black to white balls is P/Q, where $P+Q=1$; that is to say with each bag is associated a proportion P, and of the L bags there are $L.F(P)$ with proportion P. For the moment I do not assume $F(P)$ to be a continuous function. The man now draws first n balls and then m out of a bag of unknown constitution (if the number of balls in each bag were not very large he would need to replace the ball after each draw); he repeats this process N times, always choosing a bag at random. If N is made very large, then in the limit $N.F(P)$ of these draws will come from "P-bags," and again in the limit he will expect to find that in

$$\frac{n!}{p!\,q!} P^p (1-P)^q NF(P)$$

of these cases the first sample contains p black and q white balls. The total number of the N draws which will give p black and q white balls will therefore tend to

$$S_P \left\{\frac{n!}{p!\,q!} P^p (1-P)^q NF(P)\right\} = M_p,$$

---

* *Biometrika*, Vol. XVI. p. 192.

where S_P means the summation for all existing constitutions of the L bags. Hence of the M_p draws which gave p and q the proportion which came from "P-bags" will tend to

$$_PC_p = \frac{P^p(1-P)^q F(P)}{S_P\{P^p(1-P)^q F(P)\}}.$$

Now consider the second sample of m balls. We have:

Number of draws which gave p and q in first sample $\rightarrow M_p$.

Number of these which came from "P-bags" $\rightarrow M_p \, . \, _PC_p$.

Number of these last draws which will give r black and s white in the second sample $\rightarrow \frac{m!}{r!\,s!} P^r(1-P)^s M_{p\,P}C_p$.

Hence confining our attention to the M_p draws, the number which will give r and s in the second sample will tend to

$$S_P\left\{\frac{m!}{r!\,s!} P^r(1-P)^s M_{p\,P}C_p\right\} = M_{p,r},$$

and the "chance" of getting r and s after p and q, or the value to which the proportion of times that this will happen will tend as N is increased, is seen to be·

$$\frac{M_{p,r}}{M_p} = \frac{m!}{r!\,s!}\,\frac{S_P\{P^{p+r}(1-P)^{q+s}F(P)\}}{S_P\{P^p(1-P)^q F(P)\}} \quad \text{..................(i } a).$$

Put in this way the problem is as definite as that of drawing from a single bag, where the distribution of the results of successive draws is given by the terms of the binomial. Instead of drawing from a single bag we are now drawing, as it were, from a bag of bags.

Consider the expression

$$S_P\{P^{p+r}(1-P)^{q+s}F(P)\},$$

and write

$$z = P^{p+r}(1-P)^{q+s}, \quad y = F(P).$$

Suppose that the number of bags with which the experimenter is surrounded be very large, and that the values of P are distributed at vanishingly small intervals of h between 0 and 1, so that there are y_0 for which $P=0$, y_1 for which $P=h, \ldots y_l$ for which $P=1$ where $hl=1$. Then if $F(P)$ is a continuous function with ordinates $y_0, y_1 \ldots y_l$ at intervals of h, we have from the Euler-Maclaurin Theorem

$$h \overset{1}{\underset{P=0}{S}}(zy) = \int_0^1 zy\,dP + \tfrac{1}{2}h(z_0y_0 + z_ly_l)$$
$$- \tfrac{1}{12}h^2\left\{\left(\frac{d(zy)}{dP}\right)_0 - \left(\frac{d(zy)}{dP}\right)_l\right\} + \tfrac{1}{720}h^4\left\{\left(\frac{d^3(zy)}{dP^3}\right)_0 - \left(\frac{d^3(zy)}{dP^3}\right)_l\right\} - \ldots,$$

where the subscripts 0 and l indicate the values of the functions or their derivatives at $P=0$ and $P=1=lh$.

Hence the relation (i a) may be written

$$C_{p,r} = \frac{m!}{r!\,s!}\,\frac{\int_0^1 P^{p+r}(1-P)^{q+s}F(P)\,dP}{\int_0^1 P^p(1-P)^q F(P)\,dP} \quad \text{......................(i),}$$

provided that $F(P)$ is a continuous function and that in the limit as $h \rightarrow 0$, the expressions

$$hzy, \quad h^2 \frac{d(zy)}{dP}, \quad h^4 \frac{d^3(zy)}{dP^3}, \ \dots$$

vanish both when $P = 0$ and $P = 1$, for all values of $p + r$ from 0 to $n + m$.

$C_{p,r}$ is then the chance of getting r and s in a sample of m after finding p and q in a sample of n, under the conditions of sampling outlined above.

We have now to consider whether this analogy of the bags of balls can be made to correspond in any way with the problems of practical statistics. What is a bag of balls, and what the frequency distribution $F(P)$? I am here making no use of Bayes' idea of an underlying variate x with a limiting value ξ; a bag of balls represents a population that we are sampling, the individuals in which are divided into two categories, those that have and those that have not the character under consideration. The proportions are as P to Q. In speaking of random sampling we mean, I think, that there is no correlation between the order of our choice (whether in space or time), and the underlying factors which determine the character or event in question. It has been argued that outside the realm of bags of balls, few or no populations exist of sufficient stability to justify the application of Bernoulli's Theorem. Yet it seems now hardly necessary to provide evidence to prove that the stability of statistical ratios is a matter of practical reality, so that cases occur again and again in which the trained statistician can feel reasonably confident that he is taking random samples from a single population. The experiments which follow will however provide some further evidence on this point.

It is when we come to consider the distribution $F(P)$ that we reach the crucial point of the whole problem. Can we find any clear correspondence between the frequency distribution of "P-bags" and the distributions and chances of experience? In reaching equation (i) I have dealt as far as possible in terms of frequencies rather than probabilities, and it will be clearest to continue in this manner. Just as in dealing with the binomial the "chance" of a single draw can only really be interpreted by reference to the results of a series of drawings, so here the expression $C_{p,r}$ is only intelligible in connection with a series of trials. These trials are not now samples from one "bag" or one population, but samples from many "bags," samples from the varied populations that may be met with in statistical experience. To reach the simple Bayes' Equation, we must suppose $F(P)$ to be constant—the frequency distribution to be a rectangle—so that*

$$C_{p,r} = \frac{m!}{r!\,s!} \frac{\int_0^1 P^{p+r}(1-P)^{q+s}\,dP}{\int_0^1 P^p(1-P)^q\,dP} = \frac{m!}{r!\,s!} \frac{B(p+r+1, q+s+1)}{B(p+1, q+1)} \quad \text{......(ii).}$$

---

* Using the previous notation, if $F(P) = c$, then $zy = cP^{p+r}(1-P)^{q+s}$, and the differentials vanish after the $n+m$th, so that there are only a finite number of terms hzy, $h^2 \frac{d(zy)}{dP}$, ... whose value can be made as small as we please at $P = 0$ and $P = 1$ by making $h \rightarrow 0$. Hence (ia) may be written as (i).

If this expression is to be of value to the statistician, he must feel confident that among the "bags" with which he is surrounded or the populations with which he is dealing, all values of P from 0 to 1 are approximately equally represented. Or it must be shown that some better and more general value of $F(P)$ should be chosen, or alternatively that variations in $F(P)$ hardly affect the value of $C_{p,r}$. This is the great stumbling-block of Bayes' Theorem; we may be better able to surmount it after considering the results of the extensive sampling to be discussed below, but first there are a few further points of theory to be considered.

3. *The Hypergeometrical Series.* The values of $C_{p,r}$, from (ii), for values of r, 0, 1, 2, ..., m, are the successive terms of the hypergeometrical series

$$C_0\left\{1+\frac{m}{1!}\frac{p+1}{q+m}+\frac{m(m-1)}{2!}\frac{(p+1)(p+2)}{(q+m)(q+m-1)}\right.$$
$$\left.+\frac{m(m-1)(m-2)}{3!}\frac{(p+1)(p+2)(p+3)}{(q+m)(q+m-1)(q+m-2)}+\ldots\right\}\ldots(\text{iii}),$$

where

$$C_0=\frac{\Gamma(n+2)\,\Gamma(q+m+1)}{\Gamma(n+m+2)\,\Gamma(q+1)}.$$

This series may be represented by a histogram whose momental constants are*

$$\left.\begin{aligned}
\text{Mean}&=\frac{m(p+1)}{n+2},\ \text{measured from the centre of first block}\\
\mu_2&=\frac{(p+1)(q+1)\,m\,(n+m+2)}{(n+2)^2(n+3)}\\
\beta_1&=\frac{(q-p)^2(n+2m+2)^2}{\mu_2(n+2)^2(n+4)^2}\\
\beta_2&=\frac{1}{\mu_2(n+4)(n+5)}\left\{(n+1)(n+2)+6m(n+m+2)\right.\\
&\quad\left.+\frac{3(p+1)(q+1)}{(n+2)^2}\left[(n+2)^2(m-2)+m^2(n+2)-6m(n+m+2)\right]\right\}
\end{aligned}\right\}\ldots(\text{iv}).$$

Subject to certain limitations†, the histogram of the hypergeometrical series (iii) can therefore be represented by a Pearson curve having the moments of (iv).

Now let us consider the effect of introducing a function $F(P)$ into (ii) as in (i). The curve chosen must lie entirely between $P=0$ and $P=1$, and we will suppose it to be symmetrical about $P=\frac{1}{2}$, so that the number of "bags" with a proportion of P black balls is equal to that with a proportion of P white balls. The function

---

* See *Phil. Mag.* 1907, Vol. XIII. p. 370, and also *Biometrika*, Vol. XVI. pp. 157—162, where the moments of the hypergeometrical series are given with somewhat different notation. In *Biometrika*, Vol. XIII. pp. 9—14, K. Pearson obtains the equation of a Type I curve corresponding to the hypergeometrical series by putting an approximate ratio of slope to ordinate in the general differential equation, but in general slightly better results are obtained by fitting from moments.

† These moments have not been corrected for grouping or abruptness, and therefore the correspondence between histogram and curve will not be satisfactory if m is very small or if there is marked abruptness at one or both of the tails.

$$y = \frac{\Gamma(2(\alpha+1))}{\{\Gamma(\alpha+1)\}^2} P^\alpha (1-P)^\alpha \text{ ....(v)}$$

will provide us with a considerable range of unimodal curves. We know that

$$\int_0^1 y\,dP = 1,$$

and further that

(*a*) for $-1 < \alpha < 0$, (v) represents a series of U-curves with infinite ordinates at $P = 0$ and 1, but a finite area underneath;

(*b*) for $0 < \alpha < +1$, the curves are inverted U's with tangents parallel to the y-axis at $P = 0$ and 1;

(*c*) for $+1 < \alpha$, the curves are of the "cocked hat" type touching the P-axis at 0 and 1.

The general form of these three types of curves is shown in Fig. 1. If the curve of (v) is to be extended to represent the distribution of "chances" among the populations of general experience, it is almost certain that one of the forms (*a*) or (*b*) will be the most appropriate. For (*c*) implies that the frequency of "chances" between 0 and some small fraction δP is vanishingly small, which certainly appears contrary to general experience.

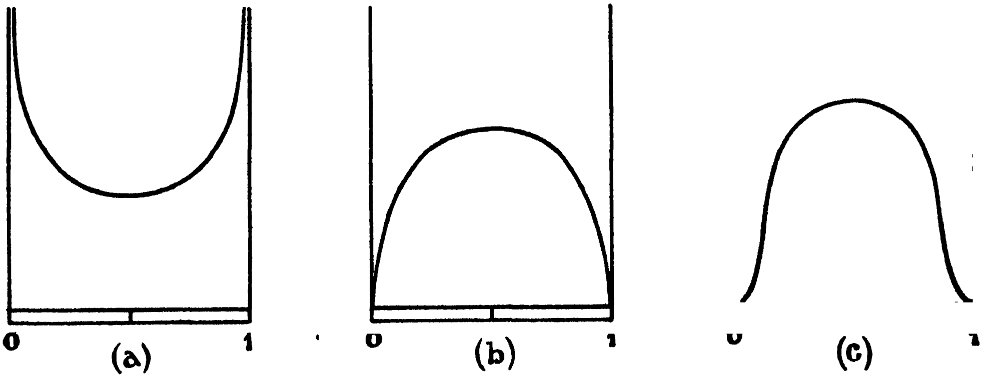


Fig. 1. Different forms of Type I Curves.

If $F(P)$ is represented by (v), or $y = cP^\alpha(1-P)^\alpha$, is it permissible to pass from (i *a*) to (i) by the Euler-Maclaurin bridge? In the first place, if $-1 < \alpha < 0$, (v) can only be appropriate if there are no populations with $P = 0$ exactly, for the proof has been developed supposing $NF(P)$ to be always finite. We must suppose that the values of P are again distributed at intervals of h, but that there are now $y_{\frac{1}{2}}$ cases with $P = \frac{1}{2}h$, $y_{\frac{3}{2}}$ with $P = \frac{3}{2}h$, ... $y_{1-\frac{1}{2}h}$ with $P = 1 - \frac{1}{2}h$, where h can be made as small as we please but never quite vanishes. Then the Euler-Maclaurin Theorem gives

$$h \overset{1-\frac{1}{2}h}{\underset{P=\frac{1}{2}h}{S}} (zy) = \int_{\frac{1}{2}h}^{1-\frac{1}{2}h} zy\,dP + \tfrac{1}{2}h\,[zy]_{\frac{1}{2}h}^{1-\frac{1}{2}h} - \tfrac{1}{12}h^2 \left[\frac{d(zy)}{dP}\right]_{\frac{1}{2}h}^{1-\frac{1}{2}h} +$$

The most unfavourable case for the vanishing of the limit terms will be when $p+r=0$, at the limit $P=\frac{1}{2}h$. Then

$$zy = cP^{a}(1-P)^{n+m+a},$$

$$\frac{d(zy)}{dP} = c\left\{aP^{a-1}(1-P)^{n+m+a} - (n+m+a)P^{a}(1-P)^{n+m+a-1}\right\} \ldots \text{ etc}$$

Putting $P=\frac{1}{2}h$, we have

$$\text{Limit } (h \to 0)\, hzy = (\tfrac{1}{2})^{a} ch^{1+a},$$

$$\text{Limit } (h \to 0)\, h^{2}\frac{d(zy)}{dP} = c\left\{a(\tfrac{1}{2})^{a-1}h^{1+a} - (n+m+a)(\tfrac{1}{2})^{a}h^{2+a}\right\}.$$

But as $-1 < a < 0$, these expressions and also those for the higher differentials can be made as small as we please by reducing h.

Now $\int_0^1 zy\,dP = c\int_0^1 P^{a}(1-P)^{n+m+a}\,dP$, which is a B-Function with a finite value; also zy is a continuous function in the range $P=\frac{1}{2}h$ to $1-\frac{1}{2}h$. Hence as h is decreased,

$$\int_{\frac{1}{2}h}^{1-\frac{1}{2}h} zy\,dP \to \int_0^1 zy\,dP \quad\text{and}\quad h \underset{P=\frac{1}{2}h}{\overset{1-\frac{1}{2}h}{S}} (zy) \to \int_0^1 zy\,dP.$$

It follows that, in general,

$$h \underset{\frac{1}{2}h}{\overset{1-\frac{1}{2}h}{S}} \left\{cP^{p+r+a}(1-P)^{q+s+a}\right\} \to \int_0^1 cP^{p+r+a}(1-P)^{q+s+a}\,dP \text{ as } h \to 0,$$

for all values of $p+r$ between 0 and $n+m$, and h being a common factor in numerator and denominator, we obtain from (i a),

$$C_{p,r} = \frac{m!}{r!\,s!}\,\frac{\int_0^1 P^{p+r+a}(1-P)^{q+s+a}\,dP}{\int_0^1 P^{p+a}(1-P)^{q+a}\,dP} = \frac{m!}{r!\,s!}\,\frac{B(p+r+a+1,\, q+s+a+1)}{B(p+a+1,\, q+a+1)} \ldots\text{(vi)}.$$

In the distribution of populations, $cP^{a}(1-P)^{a}\,\delta P$ represents the frequency of cases in which P lies between P and $P+\delta P$, but there are no populations in which P is exactly zero or unity.

From (vi) we can reach the hypergeometrical series corresponding to (iii) and the moments corresponding to (iv) by substituting

$$p+a \text{ for } p, \quad q+a \text{ for } q, \quad n+2a \text{ for } n.$$

Now if p and q are both large compared with a—that is compared with unity—the introduction of $F(P)$ will not seriously modify the form of curve given by the moments of (iv). For example, consider the change in the first two momental constants.

The Mean

$${}_{a}\nu_1' = \frac{m(p+a+1)}{n+2a+2} = \frac{m(p+1)}{(n+2)}\left(1+\frac{a}{p+1}\right)\Big/\left(1+\frac{2a}{n+2}\right)$$

$$= \nu_1'\left\{1+\frac{a}{p+1}-\frac{2a}{n+2}\right\},$$

if squares and higher powers of $\frac{a}{p+1}$ are negligible.

The Second Moment. By a similar process, it is found that

$$ {}_a\mu_2 = \mu_2\left(1 + \frac{\alpha}{p+1} + \frac{\alpha}{q+1} + \frac{2\alpha}{n+m+2} - \frac{4\alpha}{n+2} - \frac{2\alpha}{n+3}\right), $$

if the powers of $\frac{\alpha}{p+1}$ and $\frac{\alpha}{q+1}$ are neglected.

If n, the size of the first sample, is of the order of 100 or so, and neither p nor q is less, let us suppose, than 20, it is clear that the modifications in ν_1' and μ_2 will be very small. In fact they are only serious when p (or q) approaches zero, and the relations above give us a first order approximation

$$ \nu_1'\left(1 + \frac{\alpha}{p+1}\right), \quad {}_a\mu_2 = \mu_2\left(1 + \frac{\alpha}{p+1}\right). $$

It will be better to turn to the actual sampling before considering further what this result implies, and whether in fact the $F(P)$ distribution can be represented by a Type I curve.

4. *The Experimental Sampling.* The scheme of the experiment was as follows: first a certain proposition was fixed upon, e.g. taxicabs in London streets whose registration letter is LX. Then a sample of n taxis was observed in which it was noted that the proposition was true p times and untrue q times; then a second sample of m was observed in which the numbers were now r and s. After fixing upon a large number of different propositions and observing in each case the constitution of a double sample (first of n and then of m) the distribution of observed values of r could be compared with the theoretical distributions which on Bayes' Theorem the knowledge of n and p should enable us to predict. Two Series of samples were collected; in the first a very large number were obtained with a fixed size of first and second samples, viz. $n = 20$, $m = 15$; the second consists of comparatively few samples, but with values of n and m which vary from 15 to 600.

Series I. While the size of the first and second samples might be fixed at will, it was impossible to choose beforehand the size of p; that is to say, it was necessary to consider the distribution of r in second samples for all values of p from 0 to n, or rather from 0 to $\frac{n}{2}$, for we can class together cases in which the counts for the first sample give p and $n - p$ "successes*." It followed that if the number of observations for each of these values of p was to be large enough to provide a satisfactory comparison of theory with observation, the value of n selected must not be too large, or the labour involved in collecting the material would have become prohibitive. The values chosen were $n=20$, $m=15$, which provided 11 distributions for r in the second samples corresponding to $p=0, 1, 2 \ldots 10$. 12,448 double samples of 20 and 15 were collected, representing 435,680 single observations, and giving an average frequency of the order of 1000 values of r for each of the 11 values of p. The samples were taken from a very wide field, partly because this made it easier to obtain them in large numbers, but also because evidence as to the stability of statistical ratios obtained from material of very great variety would be of more value than that obtained from some limited field of inquiry. The following outline will give an idea of the main sources of the data:

---

* Whether we suppose we have found p out of n or $n - p$ out of n in the first sample, the frequency curve for the distribution of r in the second sample of m will be identical but reversed; e.g. the chance of finding 6 out of 15 after finding 8 out of 20, is the same as that of obtaining 9 out of 15 after 12 out of 20. Hence we may always take p as the size of the smaller class into which the sample of n is divided, and in second samples have only to deal (on the Bayes' hypothesis) with curves having positive skewness.

(1) *Casual observations in London streets and elsewhere.* For example:

(*a*) I walk down the Euston Road and count how many men of the first 20 I meet are smoking a pipe ($=p$), and then how many out of a succeeding 15 ($=r$).

(*b*) From a window in Gower Street I observe how many vehicles out of the first 20 that pass below are drawn by horses, and then how many out of a later sample of 15.

In taking these samples, the second sample was not necessarily observed immediately after the first, but of course it was necessary to be careful not to allow any obvious change in population constitution to occur in the interval. For instance the proportion of taxis to all motors is likely to be considerably higher in the Strand than in the Old Kent Road, and one would not take a sample of n from the first and m from the second district. But in all practical sampling this element of uncertainty arises; risks must sometimes be taken, and it was part of the present inquiry to find out how far in the long run the theoretical predictions were modified by such changes in the sampled population that a reasonable degree of caution cannot prevent sometimes slipping through.

(2) *Hackney Stud Books, Guernsey Herd Books, Greyhound Stud Books.* To ensure randomness in sampling, use was made of the alphabetical arrangements of the animals according to name; p might here be in one case the number of chestnut colts born of bay mares in a first sample of 20, and r the same thing in a sample of 15. Or the proposition might be hounds with fawn in their coats.

(3) *Letters and words in books; a great range and variety.* For example:

(*a*) p is the number of times in which the second word in the top line of the first 20 pages of Carlyle's "Latter Day Pamphlets" is a verb; r the same for the next 15 pages.

(*b*) p is the number of times in which the last word on each of 20 successive pages of the "Faerie Queene" contains the letter "e"; r is the same thing for 15 other pages.

In these samples we are testing the frequency of a certain character in the English language, or it may be the language as modified in the style of different authors.

(4) *The Census and the Registrar General's Annual Reports for England and Wales.* Here also there were a great variety of possible methods of sampling from the different tables. For example:

(*a*) Registrar General's Annual Report, 1920, pp. 126—139. The number of cases in which female deaths exceeded male deaths in samples of towns or districts chosen randomly by help of the alphabetical arrangement of names. 28 double samples could be obtained from each group of towns, one pair for each of the 28 age-at-death groups.

(*b*) *Ibid.* pp. 36—39. The number of instances in which deaths from the specified causes were above or below certain arbitrarily chosen limits, the samples of towns again usually chosen by alphabetical order.

In selecting the towns in this way it is true that we are not sampling from anything approaching an infinite population, and if town A has appeared in the first sample it does not appear in the second. This is frequently the case in practical sampling, and I do not think it vitiates the result; we may perhaps look on the two samples as drawn from the hypothetical population of an immensely greater Britain subjected to the same general conditions. We are only using the same material for a single double sample, and do not fall into the error which arises when *repeated* samples are taken without proper replacement from a limited population whose *known* constitution is made use of in calculating the theoretical frequencies to be expected in random sampling.

(5) *Information regarding Births from "The Times" of* 1895*. Here the characters observed were the proportion of male to female births, the frequency of twins, the number of days' interval between birth and announcement in the paper, etc.

The results of the 12,448 double samples classified according to the values of p and r are shown in Table I†. It will be well first to obtain from these some idea of the form of distribution of proportions or chances, P, among the populations that have been sampled.

Suppose that this distribution is represented by $F(P)$. Then on the assumption that the two samples have been drawn at random from the same population, the ratio $(p+r)/(n+m)$ or $\frac{1}{35}(p+r)$ will certainly give a likely value for P. How far will the distribution of this ratio correspond to the distribution of $F(P)$? In drawing a large number, N, of double samples we shall take a number which approaches in the limit $N.F(P)$ from "P-bags" or populations containing a proportion P of marked individuals. Of these in the limit,

$$NF(P)\frac{n+m\,!}{a\,!\,n+m-a\,!}P^a(1-P)^{n+m-a}$$

---

* Data collected for another purpose by Mr P. F. Everitt.

† The distribution of r for $p=10$ is made symmetrical, since it depended too much on the whim of the observer whether an observation $p=10$, $r=6$ (say) was recorded as (10, 6) or (10, 9), for any deductions to be drawn from the symmetry or asymmetry of the recorded distribution.

TABLE I.

Observed Distribution of p and r among the 12,448 *double samples.*

Values of p.

Values of r.	0	1	2	3	4	5	6	7	8	9	10
0	780	453	289	135	62	47	17	11	1	0	1
1	315	436	343	250	173	115	68	34	12	10	2
2	126	255	311	289	231	183	143	77	40	27	13
3	46	153	202	220	241	193	176	115	88	55	30
4	16	47	128	163	194	194	174	150	164	102	73
5	11	27	50	103	118	196	219	193	185	130	93
6	1	11	22	48	92	124	140	178	173	135	120
7	—	4	5	11	49	78	110	135	148	153	172
8	—	—	4	8	31	47	62	95	100	131	[172]
9	—	—	3	1	12	20	35	37	76	90	[120]
10	—	—	1	1	3	14	17	21	60	63	[93]
11	—	—	1	—	4	3	11	14	27	36	[73]
12	—	—	—	—	1	1	2	3	8	16	[30]
13	—	—	—	—	—	—	—	—	4	1	[13]
14	—	—	—	—	—	—	—	—	1	6	[2]
15	—	—	—	—	—	—	—	—	—	—	[1]
Totals	1295	1356	1359	1229	1211	1215	1174	1063	1087	955	[1008]

will contain a marked individuals. Hence it follows that in the N double samples the number in which $p + r = a$ will tend to

$$L_a = S_P \left\{ NF(P) \frac{(n+m)!}{(a)!\,(n+m-a)!} P^a (1-P)^{n+m-a} \right\}$$

$$= \int_0^1 \frac{(n+m)!}{(a)!\,(n+m-a)!} NF(P)\, P^a (1-P)^{n+m-a}\, dP \quad \text{.........(vii)},$$

if $F(P)$ may be taken as continuous and the conditions of p. 392 are satisfied. Consider different forms of $F(P)$.

(a) Suppose all values of the proportions among the populations sampled are equally common, or $F(P) = 1$.

Then
$$L_a = N \int_0^1 \frac{(n+m)!}{(a)!\,(n+m-a)!} P^a (1-P)^{n+m-a}\, dP$$

$$= N \frac{(n+m)!}{(a)!\,(n+m-a)!} \frac{\Gamma(a+1)\,\Gamma(n+m-a+1)}{\Gamma(n+m+2)}$$

$$= \frac{N}{n+m+1}.$$

That is to say in N samples the observed values of $p + r$ will tend to be distributed uniformly among the 36 values, 0, 1, 2 ... 35.

(b) Suppose
$$F(P) = \frac{\Gamma(2(\alpha+1))}{\{\Gamma(\alpha+1)\}^2} P^\alpha (1-P)^\alpha \quad \text{.................(v) } bis.$$

Then
$$L_a = N \frac{(n+m)!}{(a)!\,(n+m-a)!} \frac{\Gamma(2(\alpha+1))}{\{\Gamma(\alpha+1)\}^2} \frac{\Gamma(\alpha+a+1)\,\Gamma(\alpha+n+m-a+1)}{\Gamma(n+m+2\alpha+2)} \text{...(viii)}.$$

The frequency distribution of $(p+r)/(m+n)$ or L_a is therefore represented by the terms of a hypergeometrical series which may be written

$$C_0 \Bigl(1 + \frac{n+m}{1!}\,\frac{\alpha+1}{n+m+\alpha} + \frac{(n+m)(n+m-1)}{2!}\,\frac{(\alpha+1)(\alpha+2)}{(n+m+\alpha)(n+m+\alpha-1)} + \dots \qquad \text{......(ix).}$$

Now K. Pearson has shown*, from a consideration of the ratio of slope to ordinate, that the hypergeometrical series (iii) of p. 393 above can be represented by a Type I curve

$$y = y_0 \left(1 - \frac{x}{b_1}\right)^{s_1} \left(1 + \frac{x}{b_2}\right)^{s_2} \qquad \text{..........................(x),}$$

where after a slight modification of notation, and taking c the interval between the blocks of the hypergeometrical as unity,

$$s_1 = \tfrac{1}{2} n \left\{1 + \frac{1}{b}\,\epsilon\,(q-p)\right\}, \quad s_2 = \tfrac{1}{2} n \left\{1 - \frac{1}{b}\,\epsilon\,(q-p)\right\},$$

$$b_1 = \tfrac{1}{2} \{b + \epsilon\,(q-p)\}, \quad b_2 = \tfrac{1}{2} \{b - \epsilon\,(q-p)\},$$

$$\epsilon = \frac{n + 2m + 2}{2n}, \quad b = \sqrt{\epsilon^2 n^2 - pq}.$$

Now (iii) becomes (ix) on writing α for p and q, 2α for n, and $n+m$ for m, whence

$$\epsilon = \frac{n+m+\alpha+1}{2\alpha} \quad \text{and} \quad s_1 = s_2 = \alpha,$$

$$b = \sqrt{(n+m+\alpha+1)^2 - \alpha^2}, \quad b_1 = b_2 = \tfrac{1}{2} b.$$

If α is small compared to $n+m+1$, b the range of the curve approaches $n+m+1$, so that (x) may be written

$$y = y_0 \left(1 - \frac{x}{\tfrac{1}{2}(n+m+1)}\right)^{\alpha} \left(1 + \frac{x}{\tfrac{1}{2}(n+m+1)}\right)^{\alpha} \qquad \text{............(xi),}$$

where the origin is at the centre or mean. That is to say the form of distribution of the values of $p+r$ tends to follow a Type I curve with range lying between $-\tfrac{1}{2}$ and $n+m+\tfrac{1}{2}$, of precisely the same form as (v) *bis*, the distribution of population proportions.

(*c*) If the distribution of P can be represented by a high-order parabola

$$y = c_0 + c_1 P + c_2 P^2 + \dots + c_n P^n \qquad \text{........................(xii),}$$

where the constants c_0, c_1 ... form a decreasing series, it can again be shown that the distribution of $\dfrac{p+r}{m+n}$ observed in samples tends to that of $F(P)$†.

---

* *Biometrika*, Vol. XIII. p. 10 *et seq.*

† In this case $$L_a = \frac{N}{n+m+1} \left(c_0 + c_1 \frac{a+1}{n+m+2} + c_2 \frac{(a+1)(a+2)}{(n+m+2)(n+m+3)} + \dots\right).$$
If the constants c_0, c_1 ... decrease fairly rapidly this expression approximates but does not exactly correspond to the form

$$y = \frac{N}{n+m+1} \left(c_0 + c_1 \frac{a}{n+m} + c_2 \frac{a^2}{(n+m)^2} + \dots\right),$$

which we should have if the distribution $\dfrac{p+r}{n+m}$ were a similar curve to (xii). The discrepancy will be greatest for small values of $a = p+r$.

We shall therefore expect that in general the distribution of $(p+r)/(n+m)$ will give a fair approximation to the distribution of proportions or values of P in the populations sampled, the difference being greatest for small values of $p+r$.

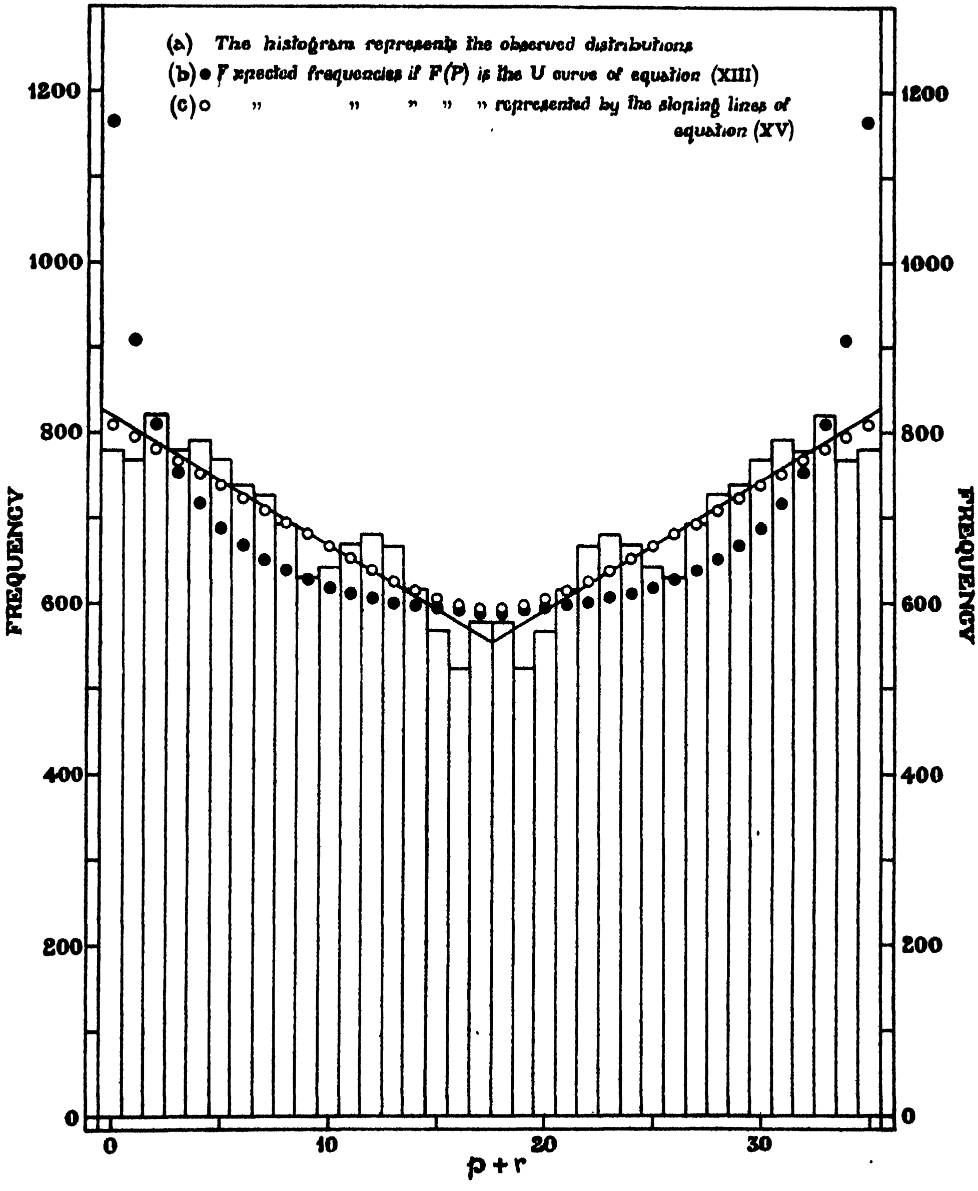


Fig. 2. Distribution of Frequencies of $p+r$ in 12,448 samples of 35 (made symmetrical).

The actual distribution observed in the present experiment is shown in Fig. 2 and the frequencies in Table II. They have been made symmetrical about

$(p+r)/(n+m)=\frac{1}{2}$, for asymmetry would have no meaning*. It will be seen at once that the distribution is not a horizontal straight line, and that the deviations from linearity are far greater and more systematic than could be expected to arise in random sampling from a series of populations in which the values of P were distributed equally between 0 and 1. Now the material is in a sense composite, being collected from various distinct sources; the average value of $p+r$ from source (5) was very low, while that from (4) was probably slightly higher than that from (3). At the same time in collecting the material there was no doubt a half-conscious effort to choose the discriminating characters so as to spread out the values of p between 0 and 10. It cannot therefore be claimed without other evidence that the distribution of Fig. 2 represents the distribution of proportions that other statisticians might find in their practical experience. For the moment we must be content with putting this question: Given that the distribution of proportions among the populations sampled was in this particular series of experiments such as to cause the distribution of $p+r$ observed, how far are the predictions of the simple form or a modified form of Bayes' Theorem borne out?

TABLE II.

Frequencies of Observed and Theoretical Values of $p+r$.

$p+r$	Observed Frequency	Theory (b)	Theory (c)	$p+r$	Observed Frequency	Theory (b)	Theory (c)
0	780	1166	809	9	630	628	681
1	768	908	795	10	643	619	667
2	821	811	781	11	670	612	653
3	779	754	767	12	682	606	639
4	792	717	752	13	668	601	626
5	769	689	738	14	616	598	615
6	739	668	724	15	568	595	605
7	727	652	709	16	524	593	598
8	694	639	695	17	578	592	594
				Totals	12,448	12,448	12,448

5. *Analysis of results.* We may proceed in various ways.

(a) Let us first compare the observed distributions of r in Table I with the theoretical distributions of r for given values of p calculated from the simple form of Bayes' Theorem, the $C_{p,r}$ of Equation (ii). The theoretical frequencies are given in Table III. If the χ^2 Test for Goodness of Fit is applied to the corresponding columns of Tables I and III, we obtain the results given in Table IV†. There is great variety in the goodness of fit, the average value of $P(\chi^2)$ being ·228. We can reach a measure of the closeness of correspondence of the whole of Table I to

---

* In a series of propositions chosen completely at random we might expect theoretically to find symmetry, but in actual practice for convenience in counting the group frequency entered as p is usually the smaller of the two. For example if I find 1 bearded man and 19 beardless I shall enter p in the records as 1 rather than 19, for it will have been the beards that I have counted.

† Small frequencies were clubbed together so as to give no groups in the theoretical frequencies containing less than 10 observations.

Table III as follows. To compare the goodness of fit of N observed to N theoretical frequencies when there are j linear relations that must hold among the frequencies, we must enter the Tables of Goodness of Fit with X^2 and N', where

$$X^2 \text{ is the sum of all } \chi^2\text{'s},$$
$$N' = N - j + 1.$$

In the present case $X^2 = 150{\cdot}1603$, $N = 103$, $j = 11$, $N' = 93$*. Whence using the Tables of the Incomplete Γ-Function it is found that $P(\chi^2) = {\cdot}00014$. That is to say the chance that the differences between Tables I and III taken as a whole are due to random sampling is extraordinarily small. This disagreement is not perhaps surprising if we remember how far the $p + r$ distribution of Fig. 2 differs from a horizontal straight line.

TABLE III.

Theoretical Distribution of p and r among 12,448 *double samples deduced from Bayes' Theorem unmodified.*

Values of p.

Values of r.	0	1	2	3	4	5	6	7	8	9	10
0	755·4	452·0	253·1	124·9	65·4	33·8	16·4	7·1	3·3	—	—
1	323·7	398·8	345·2	234·1	158·1	101·5	59·2	30·6	17·0	8·9	4·5
2	133·3	253·8	302·0	264·3	221·4	171·6	118·4	71·5	45·7	23·5	13·7
3	52·5	137·5	211·1	229·1	231·6	212·4	171·1	119·1	87·1	51·0	33·5
4	19·7	66·5	126·7	165·9	198·5	212·4	197·4	157·3	130·7	86·5	63·9
5	10·4	29·3	67·3	104·3	145·6	179·7	191·1	173·0	162·5	121·1	100·4
6	—	11·8	32·0	57·9	93·3	131·8	159·2	163·0	172·3	144·2	133·9
7	—	6·3	13·8	28·6	52·8	84·7	115·7	133·3	158·3	148·3	154·1
8	—	—	7·8	12·6	26·4	47·9	73·6	95·2	126·6	132·7	[154·1]
9	—		—	7·3	11·6	23·7	40·9	59·3	88·1	103·2	[133·9]
10	—	—	—	—	6·3	10·2	19·6	31·8	52·9	69·2	[100·4]
11	—			—		5·3	8·0	14·5	26·9	39·3	[63·9]
12	—		—	—	—	—	3·4	7·3	11·1	18·4	[33·5]
13	—	—	—	—	—	—	—	—	4·5	8·7	[13·7]
14		—	—	—	—	—	—	—	—	—	[4·5]
15	—	—	—	—	—	—	—	—	—	—	—
Totals	1295·0	1356·0	1359·0	1229·0	1211·0	1215·0	1174·0	1063·0	1087·0	955·0	[1008·0]

Where small tail frequencies have been grouped together, these figures are put below (or above) horizontal bars.

(*b*) Take now the general form of $C_{p,r}$ of Equation (i) and put for $F(P)$,

$$y = \frac{\Gamma(2(\alpha+1))}{\{\Gamma(\alpha+1)\}^2} P^\alpha (1-P)^\alpha \quad \text{....................(v) } \textit{bis}.$$

It has been shown above that an appropriate value for α can be obtained by fitting a Type I curve to the observed values of $p + r$ (Table II). Fitting by

---

* $j = 11$, corresponding to the 11 columns, $p = 0, 1, 2, \ldots 10$, for each of which the totals of theory and observation have been made to agree.

TABLE IV.

Goodness of Fit Tests. Observation and simple Bayes' Theorem.

p	χ^2	n'	$P(\chi^2)$
0	3·1806	6	·673
1	8·3146	7	·215
2	16·0261	8	·025
3	22·0819	9	·005
4	9·0518	10	·433
5	14·5416	11	·149
6	19·2740	12	·056
7	19·1419	11	·038
8	24·2553	12	·012
9	8·2975	11	·600
10	5·9950	6	·307
Totals	150·1603	103	Average ·228

equating the 2nd and 4th moments but using no corrections for abruptness I find the equation referred to 17·5 as origin is

$$y = y_0\left(1 + \frac{x}{17{\cdot}35}\right)^{-{\cdot}2266}\left(1 - \frac{x}{17{\cdot}35}\right)^{-{\cdot}2266} \quad \ldots\ldots\ldots\ldots\ldots(\text{xiii}).$$

The range of this curve is rather too short, and the values of the constants could certainly be bettered by using abruptness coefficients. If the ends of the curve are fixed at $-17{\cdot}5$ and $+17{\cdot}5$, and α found from the 2nd moment only, we have

$$y = y_0\left(1 + \frac{x}{17{\cdot}5}\right)^{-{\cdot}2039}\left(1 - \frac{x}{17{\cdot}5}\right)^{-{\cdot}2039} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(\text{xiv}).$$

Taking α as $-{\cdot}2266$ and making use of (viii), we can obtain the "expected" frequency distribution of $p + r$. This is given in the 3rd column of Table II and represented by black circles in Fig. 2. The correspondence with observation is not satisfactory at the ends; in fact a U-shaped curve does not provide a good fit because the frequencies of $p + r$ appear to be almost constant between 0 and 5. This feature will be referred to again in the second experiment.

(*c*) A more satisfactory fit to the $p + r$ distribution can be obtained from two sloping straight lines. Assuming that this distribution may be taken as an approximation to $F(P)$, I obtained by the method of least squares the lines*

$$\left.\begin{aligned} y &= 1{\cdot}191\,182 - {\cdot}764\,728P \quad &&\text{for } P = 0 \text{ to } \tfrac{1}{2} \\ y &= 1{\cdot}191\,182 - {\cdot}764\,728\,(1 - P) \quad &&\text{for } P = \tfrac{1}{2} \text{ to } 1 \end{aligned}\right\} \ldots\ldots\ldots\ldots(\text{xv}),$$

which for short may be written as

$$y = \alpha - \beta P, \quad y = \alpha - \beta\,(1 - P).$$

It is now necessary to insert this form of $F(P)$,

---

* The lines shown in Fig. 2 are those obtained directly from least squares, i.e. $y = 819{\cdot}29 - 15{\cdot}0279x$ and $y = 819{\cdot}29 - 15{\cdot}0279\,(85 - x)$ referred to $p + r = 0$ as origin. The P distribution is that of $x/85$, multiplied by a constant to make the area under the lines between $P = 0$ and 1 equal to unity.

(i) into the expression L_a of (vii), in order to get back for purposes of comparison to the frequency distribution of $p + r$;

(ii) into $C_{p,r}$ of (i), to obtain the theoretical frequency curves for the values of r in the second samples.

L_a becomes

$$N \frac{(n+m)!}{(a)!\,(n+m-a)!} \left\{ \int_0^{\frac{1}{2}} P^a (1-P)^{n+m-a} (\alpha - \beta P)\, dP \right.$$

$$\left. + \int_{\frac{1}{2}}^1 P^a (1-P)^{n+m-a} (\alpha - \beta(1-P))\, dP \right\}$$

$$= N \frac{(n+m)!}{(a)!\,(b)!} \left\{ \alpha \int_0^1 P^a (1-P)^b\, dP - \beta \left[\int_0^{\frac{1}{2}} P^{a+1} (1-P)^b\, dP + \int_{\frac{1}{2}}^1 P^a (1-P)^{b+1}\, dP \right] \right\}$$

$$\text{(writing } n + m - a = b)$$

$$= N \frac{(n+m)!}{(a)!\,(b)!} \left\{ \alpha B(a+1, b+1) - \beta B(a+2, b+1) \right.$$

$$\left. - \beta \left[\int_{\frac{1}{2}}^1 P^a (1-P)^{b+1}\, dP - \int_{\frac{1}{2}}^1 P^{a+1} (1-P)^b\, dP \right] \right\}.$$

Now $$\int_{\frac{1}{2}}^1 P^{a+1} (1-P)^b\, dP = \left[- \frac{P^{a+1}(1-P)^{b+1}}{b+1} \right]_{\frac{1}{2}}^1 + \frac{a+1}{b+1} \int_{\frac{1}{2}}^1 P^a (1-P)^{b+1}\, dP$$

$$= (\tfrac{1}{2})^{n+m+2} \frac{1}{b+1} + \frac{a+1}{b+1} \int_{\frac{1}{2}}^1 P^a (1-P)^{b+1}\, dP.$$

Hence substituting the factorial values of the B-functions,

$$L_a = N \left\{ \frac{\alpha}{n+m+1} - \frac{\beta(a+1)}{(n+m+1)(n+m+2)} \right.$$

$$\left. + \beta \frac{(n+m)!}{(a)!\,(b)!} \left[(\tfrac{1}{2})^{n+m+2} \frac{1}{b+1} + \frac{a-b}{b+1} \int_{\frac{1}{2}}^1 P^a (1-P)^{b+1}\, dP \right] \right\}.$$

Now $$\int_{\frac{1}{2}}^1 P^a (1-P)^{b+1}\, dP = (\tfrac{1}{2})^{n+m+2} \frac{1}{b+2} + \frac{a}{b+2} \int_{\frac{1}{2}}^1 P^{a-1} (1-P)^{b+2}\, dP,$$

and on continuing to integrate by parts becomes finally

$$(\tfrac{1}{2})^{n+m+2} \frac{1}{b+2} \left\{ 1 + \frac{a}{b+3} + \frac{a(a-1)}{(b+3)(b+4)} + \ldots + \frac{a(a-1)\ldots 2.1}{(b+3)(b+4)\ldots(n+m+2)} \right\}$$

$$= (\tfrac{1}{2})^{n+m+2} \frac{1}{b+2} A_{n+m,a}, \text{ let us say.}$$

And so finally

$$L_a = N \left\{ \frac{\alpha}{n+m+1} - \frac{\beta(a+1)}{(n+m+1)(n+m+2)} \right.$$

$$\left. + \beta (\tfrac{1}{2})^{n+m+2} \frac{(n+m)!}{(a)!\,(b+1)!} \left[1 + \frac{a-b}{b+2} A_{n+m,a} \right] \right\} \quad \ldots\ldots\ldots\text{(xvi)}.$$

The series for $A_{n+m,a}$ was computed numerically for $n + m = 35$ and $a = 0, 1, 2 \ldots 17$. The corresponding values of L_a are given in the 4th column of Table II and have been plotted in Fig. 2 as open circles. They agree far more closely with the observed values than those obtained from the U-curve, although the correspondence is not very exact. Applying the χ^2 Test for Goodness of Fit, it is found that $P(\chi^2) = \cdot 02$, or the chance is still much against the divergences being simply due to random sampling.

Inserting the value of $F(P)$ into (i), we have*

$$C_{p,r} = \frac{m!}{r!\,s!}\,\frac{\int_0^{\frac{1}{2}} P^{p+r}(1-P)^{q+s}(\alpha-\beta P)\,dP + \int_{\frac{1}{2}}^{1} P^{p+r}(1-P)^{q+s}(\alpha-\beta(1-P))\,dP}{\int_0^{\frac{1}{2}} P^{p}(1-P)^{q}(\alpha-\beta P)\,dP + \int_{\frac{1}{2}}^{1} P^{p}(1-P)^{q}(\alpha-\beta(1-P))\,dP}.$$

For a given value of p the denominator will be constant, and we have therefore to consider the changes with r in

$$\frac{m!}{r!\,s!}\left\{\alpha\int_0^1 P^{p+r}(1-P)^{q+s}\,dP - \beta\left[\int_0^{\frac{1}{2}} P^{p+r+1}(1-P)^{q+s} + \int_{\frac{1}{2}}^{1} P^{p+r}(1-P)^{q+s+1}\,dP\right]\right\},$$

which can be reduced to

$$\frac{m!}{(n+m)!}\cdot\frac{(p+r)!\,(q+s)!}{r!\,s!}\cdot\frac{L_{p+r}}{N},$$

where L_{p+r} is the expression of (xvi) with $p+r$ written for a. Thus for a given value of p,

$$C_{p,r} \propto \frac{(p+r)!\,(q+s)!}{r!\,s!}\,L_{p+r} \quad\text{.......................(xvii)},$$

and as L_a has been calculated for all values of a from 0 to 17 (and consequently also for $a = 18$ to 35) all the required values of $C_{p,r}$ can be readily computed. These provide the theoretical distributions for frequencies of r in second samples given in Table V. Testing the goodness of fit of the observed values of Table I

TABLE V. *Theoretical Distribution of p and r among 12,448 double samples on assumption that $F(P)$ is represented by two sloping straight lines.*

Values of p.

Values of r.	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
0	764·7	463·3	263·0	131·6	69·9	36·7	18·0	8·0	3·8	—	—
1	321·9	401·5	352·1	242·1	165·9	108·1	63·9	33·5	18·7	9·9	4·9
2	130·2	250·8	302·3	268·1	227·7	179·0	125·2	76·5	49·2	25·3	14·5
3	50·3	133·3	207·3	227·9	233·5	217·1	177·2	124·8	91·8	53·6	34·7
4	18·5	63·3	122·0	161·7	196·1	212·7	200·2	161·3	134·8	89·1	65·1
5	9·4	27·3	63·5	99·6	140·8	176·2	189·7	173·7	164·2	122·5	100·6
6	—	10·8	29·6	54·2	88·4	126·5	154·8	160·3	171·0	143·5	132·6
7		5·7	12·4	26·3	49·0	79·6	110·2	128·8	154·6	145·9	151·6
8	—	—	6·8	11·3	24·0	44·1	68·8	90·5	122·2	129·7	[151·6]
9	—	—	—	6·2	10·3	21·4	37·6	55·7	84·5	100·9	[132·6]
10	—	—	—	—	5·4	9·0	17·9	29·7	50·7	68·1	[100·6]
11	—	-	—	—	—	4·6	10·5	13·5	25·9	39·1	[65·1]
12	—	—	—	—	—	—	-	6·7	10·9	18·6	[34·7]
13	—	—	—	—	—	—	—	—	4·7	8·8	[14·5]
14	—	—	—	—	—	—	—	—	—	—	[4·9]
15	—	—	—	—	—	—	—	—	—	—	—
Totals	1295·0	1356·0	1359·0	1229·0	1211·0	1215·0	1174·0	1063·0	1087·0	955·0	[1008·0]

Where small tail frequencies have been grouped together, these figures are put below (or above) horizontal bars.

---

* No difficulty arises here in transforming (i *a*) into (i).

with these theoretical frequencies, we obtain the values of X^2 and $P(\chi^2)$ given in Table VI. The average value of $P(\chi^2)$ is here ·338, and the general fitting is far more satisfactory than was found when assuming the distribution of chances among the sampled populations to be rectangular. There are only two of the observed distributions which give a really bad fit, those for $p = 3$ and 8, and taken separately probably none would be discarded as entirely unsatisfactory. As a group however the run of values of $P(\chi^2)$ is rather low, and if we apply the combined test (as on p. 403) we find

$$X^2 = 118{\cdot}1350, \quad N = 103, \quad j = 11, \quad N' = 93,$$

TABLE VI.

Goodness of Fit Tests. Observation and Theory when $F(P)$ is taken as two sloping lines.

p	X^2	n'	$P(\chi^2)$
0	·9575	5	·915
1	7·5981	7	·268
2	9·7163	8	·205
3	15·1996	9	·055
4	8·5659	10	·477
5	11·7523	11	·302
6	13·7196	12	·249
7	15·8142	11	·102
8	22·2513	12	·022
9	5·8525	12	·883
10	6·7077	6	·243
Totals	118·1350	103	Average ·338

whence (using the Tables of the Incomplete Γ-Function) it is found that $P(\chi^2) = {\cdot}0346$. The odds are still much against the difference between observation and theory arising simply from random sampling, but they are not weighted nearly so heavily as in the previous case when $P(\chi^2) = {\cdot}00014$. By fitting a more complex form of curve to the frequency distribution of $p + r$, we could certainly improve the agreement to some extent, but as it cannot be claimed that the exact form of this distribution with its peculiar sinuosities has a wider significance, the labour involved would not be worth while.

(*d*) There is however another method by which we may consider the results. Returning to the analogy of the bags of balls, we expected in the long run after taking a very large number, N, of random double samples, to find M_p cases in which the first sample contained p black balls, and of these a proportion tending to $C_{p,r}$ (of Equation (i)) in which the second sample of m contained r black balls, where

$$M_p = N \int_0^1 \frac{n!}{p!\,q!} F(P)\, P^p (1-P)^q \, dP.$$

The expected frequencies of cases in which $p + r$ has a fixed value, a, are then

$$M_0 C_{0,a},\ M_1 C_{1,a-1},\ M_2 C_{2,a-2},\ \dots\ M_a C_{a,0} \quad \text{..............(xviii)},$$

for p ranging from 0 to a, where it will be found that

$$M_t C_{t,a-t} = \frac{n!\,m!\,N}{(a-t)!\,(m-a+t)!\,t!\,(n-t)!}\int_0^1 P^a(1-P)^{m+n-a}F(P)\,dP$$

$$= \frac{1}{(a-t)!\,(m-a+t)!\,t!\,(n-t)!} \times \text{a factor independent of } t.$$

Hence the proportionate frequencies of the series (xviii) are quite independent of the form of $F(P)$, the term containing this function only appearing as a common multiplying factor*. The $a+1$ observed frequencies in Table I, for which

$$p + r = a,$$

should therefore correspond to the $a+1$ theoretical frequencies,

$$\frac{K}{0!\,(n)!\,(a)!\,(m-a)!},\quad \frac{K}{1!\,(n-1)!\,(a-1)!\,(m-a+1)!},\quad \frac{K}{2!\,(n-2)!\,(a-2)!\,(m-a+2)!},\ \text{etc.}$$

Determining K to make the sums of these $a+1$ observed and theoretical frequencies agree, the figures given in Table VII were calculated for all the possible values of p and r.

TABLE VII.

Theoretical Distribution obtained by considering the proportionate frequency of p among double samples with a given p + r.

Values of p.

Values of $p + r$.	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	Totals
0	780·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	780·0
1	329·2	438·8	—	—	—	—	—	—	—	—	—	—	—	—	—	768·0
2	144·9	413·9	262·2	—	—	—	—	—	—	—	—	—	—	—	—	821·0
3	54·1	250·0	339·2	135·7	—	—	—	—	—	—	—	—	—	—	—	779·0
4	20·7	137·6	301·8	258·7	73·2	—	—	—	—	—	—	—	—	—	—	792·0
5	7·1	64·7	204·8	283·5	172·2	36·7	—	—	—	—	—	—	—	—	—	769·0
6	2·3	27·3	118·1	236·2	231·6	105·9	17·6	—	—	—	—	—	—	—	—	739·0
7	—	11·5]	61·7	168·2	238·3	176·0	62·9	8·4	—	—	—	—	—	—	—	727·0
8	—	4·0]	28·0	101·0	195·0	208·0	120·0	34·3	3·7	—	—	—	—	—	—	694·0
9	—	—	12·0]	50·9	129·8	188·8	157·4	72·6	16·9	1·6	—	—	—	—	—	630·0
10	—	—	4·7]	25·7	84·9	163·1	185·3	123·5	46·3	[9·5	—	—	—	—	—	643·0
11	—	—	—	13·5]	50·0	124·6	186·9	169·9	92·1	28·3	[4·7	—	—	—	—	670·0
12	—	—	—	5·1]	25·5	81·6	158·6	190·3	140·6	62·5	15·8	[2·0	—	—	—	682·0
13	—	—	—	—	12·6]	45·2	112·9	175·6	171·1	103·7	38 0	[8·9	—	—	—	668·0
14	—	—	—	—	4·3]	20·6	66·2	132·4	167·4	133·9	67·0	20·3	[3·9	—	—	616·0
15		—	—		—	9·4]	33·9	87·2	141·8	147·0	97·0	40·1	[11·6	—	—	568·0
16	—	—		—	—	3·0]	15·0	50·1	104·6	139·4	119·4	65·1	22·2	[5·2	—	524·0
17	—	—	—	—	—	—	7·7]	29·7	80·3	137·7	151·4	107·1	48·2	13·5	[2·4	578·0
																12,448

Where small tail frequencies in the horizontal rows have been grouped together, these figures are put outside a square bracket.

---

* In the theory of probability

$$\frac{a!}{(a-t)!\,t!}\,\frac{(m+n-a)!}{(n-t)!\,(m-a+t)!}\Big/\frac{(m+n)!}{(m)!\,(n)!}$$

is the chance of drawing t marked individuals in a sample of n taken at random from a larger group $m+n$ which contains a marked individuals. And as we should expect, within the double samples in which $p+r=a$, the frequency of samples with different values of p from 0 to a follows this law.

Applying the χ^2 test, we find the results given in Table VIII; the average value of $P(\chi^2)$ is ·358. Here

$$X^2 = 106{\cdot}0789, \quad N = 100, \quad j = 17, \quad N' = 84,$$

from which

$$P(\chi^2) = {\cdot}045^-.$$

TABLE VIII.

Goodness of Fit Tests; applied to observation and theory among double samples with the same value of $p + r$.

$p+r$	χ^2	n'	$P(\chi^2)$
1	1·0422	2	·31
2	5·4267	3	·07
3	1·3590	4	·71
4	5·0774	5	·28
5	6·6918	5	·16
6	2·8315	6	·73
7	3·5417	6	·62
8	6·8965	6	·23
9	10·1552	7	·12
10	17·3595	7	·01
11	8·9641	7	·18
12	7·4287	7	·29
13	6·2253	7	·40
14	1·9983	7	·92
15	5·7707	7	·45
16	5·7921	7	·45
17	9·5182	7	·15
Totals	106·0789	100	Average ·358

This final test appears to avoid the introduction of the uncertain function $F(P)$, and we have obtained by it a measure of the closeness of fit of observation to theory which is but very little higher than we obtained on the assumption that $F(P)$ could be represented by two sloping straight lines. This seems to suggest that no further adjustment of $F(P)$ will materially improve the fit, so that there appears definitely to be some lack of correspondence between observation and theory, a divergence not very serious perhaps, but of the order of ·05 on the "P, χ^2" scale. This can be located more clearly by a comparison of the moments of the observed frequency distributions with the theoretical values obtained on various hypotheses.

6. Table IX contains the values of the Mean, Standard Deviation and the constants β_1 and β_2 of the frequency distributions of values of r found in samples of m following p in samples of n. They have been calculated,

(1) On the simple Bayes' hypothesis, from relations (iv).

(2) On the assumption that $F(P)$ is of form (v) with $a = -{\cdot}25$, by substituting in (iv)

$$p + a \text{ for } p,\ q + a \text{ for } q \text{ and } n + 2a \text{ for } n.$$

(3) On the assumption that $F(P)$ is represented by the two straight lines of (xv); the moments were here computed from the frequencies given in Table V.

(4) From the observed frequencies given in Table I.

No corrections for grouping have been used in any of the four cases.

First compare the series of moments calculated on the hypothesis (2)—where $F(P)$ is assumed to be a U-curve—with those obtained from the simple Bayes' hypothesis (1). It will be seen that for $p=0$ the differences are large, but as p increases the two sets of values more and more nearly agree, until when $p=3$ the differences, except in the case of the mean, are not greater than the probable errors*. It is of course the steep rise of the U-curve distribution of $F(P)$ that modifies the moments at $p=0$ so greatly. The differences in (2) are in the sense of a lower mean, a smaller standard deviation and greater skewness. It has been shown† that to a first order

$$ {}_a\nu_1' = \nu_1'\left(1+\frac{\alpha}{p+1}\right), \quad {}_a\mu_2 = \mu_2\left(1+\frac{\alpha}{p+1}\right). $$

That is to say the difference between the moments of (1) and (2) does not depend primarily on the values of either n or m. For example, if the size of the first sample were 100 instead of 20, while the second sample remains at 15, the following moments will be found for the distribution in second samples when $p=3$:

	Mean	σ	β_1	β_2
$\alpha=0$	·588	·801	2·13	5·35
$\alpha=-·25$	·554	·779	2·28	5·53

The percentage differences will be found here to be only slightly greater than between the corresponding moments in Table IX, for $n=20$ and $p=3$; the average in the one case is 4·7 %, in the other 4·2 %. Beyond $p=3$, the differences will steadily decrease.

What conclusions can be drawn from this? It shows that whether the distribution of "chances" among the populations sampled is rectangular or follows the U-curve of (v) with an α of about $-·25$, the frequency curves for values of r in second samples, as judged by the values of their first four moments, do not differ greatly if p be larger than 3 or 4. That is to say if some uncertainty exists regarding the form of $F(P)$, then the larger the size of the first sample, n, is, the smaller is the fraction of the range of possible values of p affected by this uncertainty. Further, even if a U-curve does not provide a very satisfactory fit, it appears likely that the same general rule will hold for any other curve not differing very markedly from it (as, for example, the two sloping straight lines of Equations (xv)). This may appear a vague and unsatisfactory position from the standpoint of strict analysis, but it will be seen later that it is of considerable value from the practical point of view.

---

* The probable errors given are calculated for the moments found on hypothesis (3), where the size of sample is in each case that actually observed, i.e. of the order of 1000.

† On p. 396.

TABLE IX. *Moments of Distributions in second samples on various hypotheses.*

$p=0$	Mean	σ	β_1	β_2
(1) $F(P)$, constant ...	·682	1·023	3·71	7·68
(2) $F(P)$, a U-curve ...	·523	·905	5·07	9·52
(3) $F(P)$, two sloping lines	·663 ± ·019	1·003 ± ·013	3·70	7·56
(4) Observation	·641	·989	3·66	7·11
$p=1$				
(1) $F(P)$, constant ...	1·364	1·412	1·58	4·81
(2) $F(P)$, a U-curve ...	1·221	1·349	1·85	5·18
(3) $F(P)$, two sloping lines	1·327 ± ·025	1·389 ± ·018	1·61 ± ·16	4·84 ± ·37
(4) Observation	1·322	1·345	1·36	4·38
$p=2$				
(1) $F(P)$, constant ...	2·045	1·686	·87	3·86
(2) $F(P)$, a U-curve ...	1·919	1·648	·98	4·00
(3) $F(P)$, two sloping lines	1·991 ± ·030	1·658 ± ·021	·88 ± ·10	3·84 ± ·21
(4) Observation	1·899	1·619	1·15	4·86
$p=3$				
(1) $F(P)$, constant ...	2·727	1·895	·53	3·40
(2) $F(P)$, a U-curve ...	2·616	1·872	·58	3·47
(3) $F(P)$, two sloping lines	2·656 ± ·036	1·867 ± ·025	·55 ± ·07	3·42 ± ·17
(4) Observation	2·525	1·734	·39	3·15
$p=4$				
(1) $F(P)$, constant ...	3·409	2·059	·33	3·13
(2) $F(P)$, a U-curve ...	3·314	2·046	·36	3·16
(3) $F(P)$, two sloping lines	3·323 ± ·039	2·031 ± ·028	·34 ± ·05	3·13 ± ·13
(4) Observation	3·353	2·092	·53	3·44
$p=5$				
(1) $F(P)$, constant ...	4·091	2·188	·20	2·96
(2) $F(P)$, a U-curve ...	4·012	2·183	·22	2·98
(3) $F(P)$, two sloping lines	3·994 ± ·042	2·167 ± ·030	·22 ± ·04	3·00 ± ·12
(4) Observation	3·989	2·236	·18	2·87
$p=6$				
(1) $F(P)$, constant ...	4·773	2·288	·12	2·85
(2) $F(P)$, a U-curve ...	4·709	2·289	·13	2·85
(3) $F(P)$, two sloping lines	4·670 ± ·045	2·276 ± ·032	·14 ± ·03	2·88 ± ·10
(4) Observation	4·607	2·282	·15	2·86
$p=7$				
(1) $F(P)$, constant ...	5·455	2·363	·06	2·77
(2) $F(P)$, a U-curve ...	5·407	2·368	·07	2·77
(3) $F(P)$, two sloping lines	5·354 ± ·049	2·360 ± ·035	·07 ± ·02	2·78 ± ·08
(4) Observation	5·272	2·251	·02	2·84
$p=8$				
(1) $F(P)$, constant ...	6·136	2·415	·03	2·72
(2) $F(P)$, a U-curve ...	6·105	2·423	·03	2·72
(3) $F(P)$, two sloping lines	6·057 ± ·050	2·430 ± ·035	·04 ± ·01	2·73 ± ·07
(4) Observation	6·029	2·354	·13	2·77
$p=9$				
(1) $F(P)$, constant ...	6·818	2·446	·01	2·70
(2) $F(P)$, a U-curve ...	6·802	2·456	·01	2·69
(3) $F(P)$, two sloping lines	6·772 ± ·054	2·471 ± ·038	·008 ± ·003	2·67 ± ·07
(4) Observation	6·640	2·449	·029	2·76
$p=10$				
(1) $F(P)$, constant ..	7·500	2·456	·00	2·69
(2) $F(P)$, a U-curve ...	7·500	2·466	·00	2·68
(3) $F(P)$, two sloping lines	7·500 ± ·075	2·488 ± ·053	·00	2·67 ± ·09
(4) Observation	[7·500]	2·411	[·00]	2·71

Returning again to Table IX, it is seen that the values of σ, β_1 and β_2 obtained on hypothesis (3)—that $F(P)$ is two sloping straight lines—differ only very slightly from those of the simple Bayes' Theorem, (1). The Means of (3) are less than those of (1) by a varying amount which reaches a maximum of $\cdot 103$ when $p = 6$, but has a maximum percentage value of $3\frac{1}{2}\,\%$ when $p = 0$. As might be expected from the results of the χ^2 Goodness of Fit Test, the moments of the observations, (4), agree on the whole most closely with the moments of (3). The extent and sign of the difference is summarised in Table X, where I have given the differences, (4) − (3), in terms of the probable errors of (3). We see:

TABLE X.

Deviations, Moments of observed distributions—Moments of theoretical distributions ($F(P)$ two sloping lines), in terms of probable errors of the latter.

p	Mean	σ	β_1	β_2
0	−1·15	−1·02	−(small)	−(small)
1	−0·23	−2·44	−1·6	−1·3
2	−3·05	−1·80	+2·7	+4·9
3	−3·67	−5·25	−2·2	−1·6
4	+0·77	+2·18	+3·8	+2·4
5	−0·12	+2·31	−1·0	−1·0
6	−1·41	+0·21	+0·6	−0·2
7	−1·68	−3·17	−2·6	+0·7
8	−0·55	−2·17	+9·6	+0·7
9	−2·45	−0·58	+6·8	+1·3
10	—	−1·46	—	+0·4

(*a*) The means of the observed distributions show a systematic defect (there is one exception for $p = 4$), which amounts in the average to 1·35 times the probable error.

(*b*) The observed standard deviations are also smaller than the theoretical values in all cases except three.

(*c*) The observed values of β_1 and β_2 are sometimes too large and sometimes too small, but do not appear to show any systematic divergence. In particular, for low values of p, the skewness of the observed curves is not significantly below the theoretical value[*].

(*d*) There are considerably more differences of over 2 to 3 times the probable error than we should expect to find if the variations were simply the result of random sampling.

The systematic difference in mean and standard deviation can only, I think, be in small part due to an inadequacy of the lines of Equations (xv) in representing the distribution $F(P)$ of the populations sampled, for it has been seen that the test which avoided the assumptions regarding the form of this function gave a value of $P(\chi^2)$ of only ·045. A possible cause may lie in the process of sampling. If there had been a certain degree of instability among the sampled populations, there would have been a tendency for the proportion r/m to fall somewhat nearer the modal value of $F(P)$—that is in general to have a lower value[*]—than if the constitution of the population had remained constant for both samples. This would result in a lower mean, and also a smaller standard deviation[†] in the second sample than the theory would predict.

---

[*] For $p = 0$ the values of β_1 and β_2 fall far outside the range of the tables for probable errors, but I think that the observed differences ($-\cdot 04$ for β_1 and $-\cdot 45$ for β_2) must be considerably less than the probable errors.

But if it is necessary to admit a certain amount of discordance between observation and theory, we should be guilty of a lack in sense of proportion if we were to condemn Bayes' Theorem on these grounds. A glance at the distributions of Table I shows that in a very real sense there is law and order among the figures. With the observed series of distributions changing gradually from J-shaped to symmetrical curves as p increases, it would be foolish to insist on some theoretical grounds that it was useless to attempt to give numerical expression to the prediction of future events. There is clearly regularity where some of the critics would apparently expect an irregular jumble of figures, and the preceding analysis has shown that observation and theory are not so very discordant. The greatest error in the value of a mean is only 5 °/₀ (in the two cases, $p=2$ and 3), while the variation among the observed values of r is more often than not less than that predicted by theory.

It must not of course be forgotten that the best fit has only been obtained by making use of an *a posteriori* curve, $F(P)$; we shall see more clearly, after dealing with the second series of experimental samplings, how far it is possible to proceed without this help.

7. *Series II.* The first series of experiments have shown how far, if repeated samples of a fixed size are taken, the resulting distributions of r for a given p follow the theoretical rules, and the results have been on the whole encouraging. In actual practice, however, the statistician is not dealing repeatedly with single values of n and m; indeed the same values may never recur in all his experience. The position is the same in almost all statistical problems, whether we are making use of the criterion of the simple probable error or applying the χ^2 Test for Goodness of Fit. Theory gives the result to be expected in repeated tests under the same conditions, while in practice the tests are required in a number of single cases under different conditions. In our particular problem, the statistician wishes to know whether if he applies Bayes' Theorem[‡] in some 50 or 100 different problems, the range for r that he predicts will not be exceeded in his second samples more often than theory expects. If the theory lays down that in k °/₀ of cases the number of "black balls" in the second sample of m should lie between limits $L(p, n, m)$ and $L'(p, n, m)$, then he wishes to have some confidence in

---

[*] In view of the U- or V-shaped form of this function $F(P)$.

[†] Smaller, because the lower the P of the population, the less the σ.

[‡] Assuming $F(P)$ constant or to have some other definite value determined *a priori*.

expecting that out of N problems with differing values of p, n and m which his work may bring before him, he will be correct in predicting that r lies between L and L' on $\frac{1}{100}Nk \pm \epsilon$ occasions, where ϵ is a small quantity, such that ϵ/N approaches zero as N increases. Suppose that in each of some 100 problems he calculates the range within which 99% of the r's of second samples should lie, then if he accepts these ranges for purposes of prediction will he be only in error in one or about one case out of the hundred? It appears to me that it is only by this reference to the results of a number of samples that "probability limits" and "odds for" or "against" can be given any useful interpretation, and this is the way that it is proposed to look at the problem in dealing with the second experiment. I shall make use of Bayes' Theorem, assuming $F(P)$ to be constant, to calculate in each of 300 cases (in which n and m differ) the limits within which, (a) 90 °/₀, (b) 99 °/₀, of the values of r, should fall if repeated samples were taken, then apply these limits to the single instances and so find out whether in the long run the theory plays false.

The following notes indicate the chief points observed in collecting the 300 samples:

(1) As far as was humanly possible the propositions or subjects for recording were chosen in a haphazard manner without forethought as to possible values of P. To assist in making the selection as unprejudiced as possible, all but about 10 °/₀ of the propositions were noted down during the course of a fortnight before the recording was begun, and the remaining 30 were all fixed upon before any attempt was made to reduce the data.

(2) The values of n and m varied from 15 to 600 and were combined with great variety; m was often taken larger than n*, so that the second sample was to be predicted from a smaller first sample.

(3) Particulars of twenty of the double samples are given in Table XVIII and are discussed later. They were chosen randomly from the 300, viz. their numbers in the order of record were 15, 30, 45, ..., 300.

(4) In the majority of cases n and m were taken as multiples of 10; this made no difference to the value of the experiment, but considerably eased the computing. As a result certain combinations of n and m appeared a number of times.

(5) The recording was much simplified by using one of Galton's "Pocket Recorders," several of which are preserved in the Galton Laboratory. That used was a small instrument with 5 keys each working a numbered dial, fitting the fingers and thumb of the hand, which can be conveniently carried and worked unobserved in the pocket. It was possible to take as many as three different samples at the same time.

The distribution for the 300 samples of the ratios $(p+r)/(m+n)$ is shown in Fig. 3†. It is seen again to be of U- or V-shape with the same suggestion of a maximum between the proportions ·1 and ·2 as in the previous experiment. The form of this divergence from linearity in the $F(P)$ distribution will be considered later, but we will proceed with the test on the assumption (afterwards shown to be justified) that this divergence is of no practical importance.

---

* In 129 cases.

† It has been made symmetrical about 0·5 for the same reason as before.

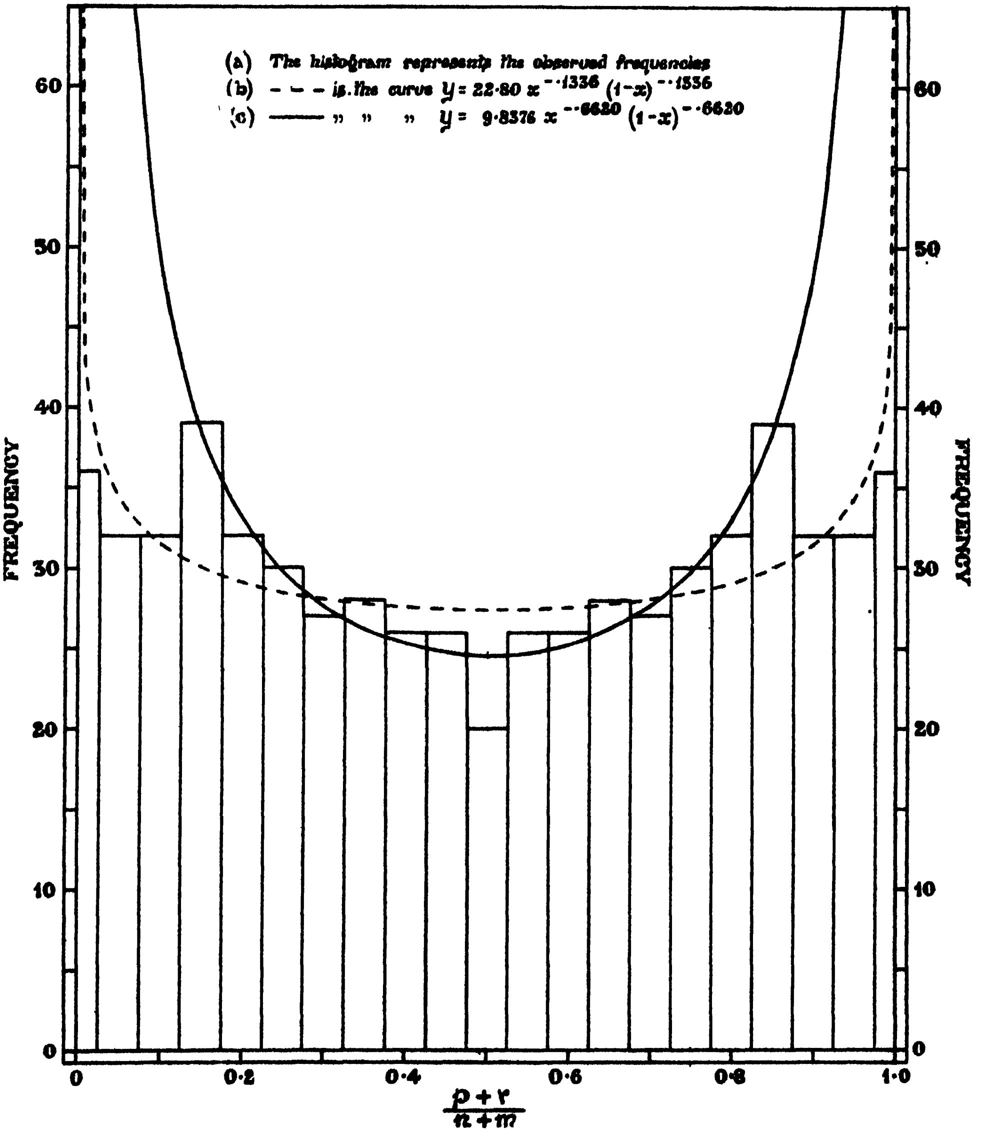


Fig. 3. Distribution of Frequencies of $\frac{p+r}{n+m}$ in 800 samples (made symmetrical).

After collecting the samples the following method of analysis was carried out. For each double sample the position of the Mean, and the values of σ, β_1 and β_2 for the frequency distribution of r in second samples were calculated from the relations (iv); this was a straightforward if very lengthy piece of computation.

For each of these frequency curves the position of the ordinates bounding what may be termed the central 90 °/₀ and central 99 °/₀ of frequency* were obtained by help of the Tables in the Appendix. These limits were chosen partly to form a test for the present experiment, but also because it seemed likely that they would be of value in other problems. Had tables of the Incomplete Beta-Function been available any percentile limits could have been chosen at will and the labour involved in calculating these four tables would have been avoided. The method by which the tables were computed is described below, but it is of some interest to consider the range which they are needed to cover. Taking the two last equations of (iv) it is found on eliminating p and q that β_1 and β_2 satisfy the linear relation,

$$\beta_2 = \frac{n+4}{n+5}\,\frac{(n+1)(n+2)+6m(m+n+2)}{(n+2m+2)^2}\,\beta_1 + \frac{n+3}{n+5}\left\{3 - \frac{2(n+2)}{m(n+m+2)}\right\} \quad \text{..............(xix).}$$

That is to say, for a given size of the two samples the point in the β_1, β_2 diagram representing the distributions in second samples lies on a straight line, its distance from the β_1 axis increasing as p/n decreases from 0·5 to 0·0. Now write $m = n\lambda$ in (xix); then

$$\beta_2 = \frac{n+4}{n+5}\,\frac{(n+1)(n+2)+6n\lambda(n(1+\lambda)+2)}{(n(1+2\lambda)+2)^2}\,\beta_1 + \frac{n+3}{n+5}\left\{3 - \frac{2(n+2)}{n\lambda(n(1+\lambda)+2)}\right\},$$

or approximately if n is not too small,

$$\beta_2 = \frac{1+6\lambda(1+\lambda)}{(1+2\lambda)^2}\,\beta_1 + 3 - \frac{2}{\lambda(1+\lambda)n} \quad \text{.................(xx).}$$

First consider the slope of this line,

$$a = \frac{1+6\lambda(1+\lambda)}{(1+2\lambda)^2}, \text{ giving } \frac{da}{d\lambda} = \frac{2}{(1+2\lambda)^3}.$$

Thus a increases with λ but at a decreasing rate; we have,

(1) $\lambda = 0$, or m is very small compared to n; then (xx) tends to

$$\beta_2 - \beta_1 - 3 + \frac{2}{m} = 0,$$

which is the line giving curves corresponding to the binomial $(p+q)^m$.

(2) $\lambda = 1$, or $m = n$; then $a = \frac{13}{9}$ or little less than $\frac{3}{2}$ the slope of the Type III line.

(3) $\lambda \to \infty$, or m is very large compared with n; then in the limit (xx) becomes the Type III line

$$2\beta_2 - 3\beta_1 - 6 = 0.$$

The constant term in (xx), or

$$b = 3 - \frac{2}{\lambda(1+\lambda)n},$$

---

* I.e. the ordinates cutting off (1) the 5·0 °/₀ tails, (2) the 0·5 °/₀ tails of the area under the frequency curve.

shows that the line cuts the β_1 axis where $\beta_2 < 3$ and the point of intersection tends to 3, the Gaussian Point, as m is made large. b will decrease as m and n decrease; taking as a lower limit for satisfactory fit of curve to histogram $n = m = 15$*, we find that (xix) cuts the β_1 axis where $\beta_2 = 2{\cdot}64$.

Thus for the Bayes' hypergeometrical series the point β_1, β_2 will fall within the Type I area and in the cases likely to be met with will probably not fall outside the area between

$$\begin{cases} 2\beta_2 - 3\beta_1 - 5 = 0, \\ 2\beta_2 - 3\beta_1 - 6 = 0. \end{cases}$$

These limits were taken into account in forming the Tables of the Appendix.

The actual distribution of values of β_1, β_2 found in the 300 samples is shown in grouped form in Table XI. The majority of points lie close to the β_1 axis but

TABLE XI. *Distribution of β_1 and β_2.*

β_2 \ β_1	0·0—·2	·2—	·4—	·6—	·8—	1·0—	1·2—	1·4—	1·6—	1·8—	2·0—	2·2—	2·4—	2·6—	2·8—	3·0—	3·2—	3·4—	3·6—	3·8—4·0	Totals
2·6—2·8	4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4
2·8—	119	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	119
3·0—	69	9	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	78
3·2—	1	24	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	26
3·4—	—	6	14	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	20
3·6—	—	—	6	7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	13
3·8—	—	—	—	5	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	7
4·0—	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2
4·2—	—	—	—	—	1	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
4·4—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
4·6—	—	—	—	—	—	1	1	1	—	—	—	—	—	—	—	—	—	—	—	—	3
4·8—	—	—	—	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	2
5·0—	—	—	—	—	—	—	—	2	2	—	—	—	—	—	—	—	—	—	—	—	4
5·2—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	1
5·4—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	1
5·6—	—	—	—	—	—	—	—	—	—	1	2	—	—	—	—	—	—	—	—	—	3
5·8—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	—	—	—	2
6·0—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	1
6·2—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6·4—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	—	—	1
6·6—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
6·8—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7·0—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	—	—	—	—	2
7·2—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7·4—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7·6—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4	—	—	—	4
7·8—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8·0—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8·2—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8·4—8·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1
Totals	193	39	21	12	5	4	1	5	2	3	3	2	—	1	—	2	4	—	—	1	298

The Table does not include the two outlying observations $\begin{cases} \beta_1 = 6{\cdot}07, \ \beta_2 = 10{\cdot}89. \\ \beta_1 = 9{\cdot}58, \ \beta_2 = 14{\cdot}35. \end{cases}$

---

* For lower values of n and m Greenwood's tables will be found very useful (*Tables for Statisticians and Biometricians*, Table XLVIII). These give the percentage frequencies of the actual terms of the hypergeometrical series for a wide range of combinations of n and m.

a long tail runs out across the sheet representing distributions of increasing skewness; these are of course the distributions in which p/n was 0, or very small. The correlation between the 298 values[*] of β_1 and β_2 is $+\cdot 9921$ and the regression line of β_1 on β_2 is

$$\beta_2 = 1\cdot 4222\,\beta_1 + 2\cdot 8289 \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xxi)},$$

a line sloping gradually away from the Type III line as β_1 increases, and drawn in Figure 5 (p. 437).

The 90 $^\circ/_\circ$ and 99 $^\circ/_\circ$ limits of r are those within which we should expect to find corresponding frequencies to fall, if a great number of series of sampling experiments were carried out. For such series of 300, the mean number of cases in which r should fall within the 90 $^\circ/_\circ$ limits is

$$300 \times \tfrac{90}{100} = 270,$$

and the variation of the number in individual series about this is given by the terms of the binomial[†], $(\frac{1}{10}+\frac{9}{10})^{300}$, or represented by a standard deviation of 5·20. Similarly for the 99 $^\circ/_\circ$ limits the mean number is 297 and the standard deviation of the binomial, 1·72.

(*a*) We may use the Tables of the Appendix. It will then be found for the actual observations that

$$\left\{\begin{array}{l}\text{271 lay within the central 90 }^\circ/_\circ\text{ limits,}\\ \text{294} \quad „ \quad „ \quad „ \quad \text{99 }^\circ/_\circ \quad „\end{array}\right.$$

The first of these numbers is in excellent agreement with theory and the second is not unexpectedly far out, for the binomial $(\frac{1}{100}+\frac{99}{100})^{300}$ is very skew and the observed number, 294, is only at a distance of 1·74 times the standard deviation from the mean, 297, in the direction of the extended tail[‡].

Consider now the result of applying two other more approximate tests for predicting the range of r in second samples:

(*b*) Using the Mean and Standard Deviation of Bayes' Theorem, but taking the curve of distribution to be always Normal. If ν_1' and σ are the first two moments calculated as in (iv),

$$\left\{\begin{array}{l}\text{the central 90 }^\circ/_\circ\text{ will lie between ordinates at } \nu_1' \pm 1\cdot 6449\,\sigma,\\ \quad „ \qquad \text{99 }^\circ/_\circ \qquad „ \qquad\qquad „ \qquad\qquad „ \quad \nu_1' \pm 2\cdot 5758\,\sigma\S.\end{array}\right.$$

---

[*] In this calculation two extreme values were omitted, viz. $\beta_1=6\cdot 07$, $\beta_2=10\cdot 89$ and $\beta_1=9\cdot 58$, $\beta_2=14\cdot 35$.

[†] As there is no correlation between the different double samples the chance of obtaining a given number of r's within the limits is the same as of drawing that number of black balls from a bag containing black to white in the ratio of (1) 90 to 10 and (2) 99 to 1.

[‡] This binomial has for moment coefficients $\beta_1=\cdot 328$, $\beta_2=3\cdot 317$, from which we find from the Tables in the Appendix that its central 90 $^\circ/_\circ$ of area lies between ordinates at $-1\cdot 46\sigma$ and $+1\cdot 80\sigma$ from the mean. Owing to the abrupt beginning of the binomial the value $-1\cdot 46\sigma$ taken from the Type I curve may not be very satisfactory, but the deviation towards the extended tail should be accurate.

[§] These are the deviates of the Normal curve corresponding to $\frac{1}{2}(1+\alpha)=\cdot 95$ and $\cdot 995$.

(*c*) Using the very crude values for Mean and Standard Deviation which are often employed, where the first sample is treated as indefinitely great,

$$ {}_0\nu_1' = \frac{mp}{n}, \quad \sigma_0 = \sqrt{\frac{mpq}{n^2}}, $$

and again taking the curve of distribution to be always Normal. Here the central areas will be limited as in case (*b*), if we substitute ${}_0\nu_1'$ for ν_1', σ_0 for σ.

These two methods of approximation have been discussed by K. Pearson in the earlier papers on the subject (*Phil. Mag.* 1907, Vol. XIII. pp. 365—378, and *Biometrika*, Vol. XIII. p. 12). The numbers of observations falling outside the different ranges are summarised in Table XII. It will be seen that there are actually fewer values of r lying outside the limits determined as in (*b*), than with the complete Bayes' Theorem constants as in (*a*). This does not of course mean that (*b*) is more satisfactory than (*a*), since for the 90 °/₀ limits the frequency 29 (*a*) is nearer the expected value 30 than the 24 (*b*); in fact by using the skew curves it is possible to predict the range within slightly narrower limits than with the normal curves, as can be seen from Tables (1) and (2) of the Appendix*. The difference between 6 and 5 cases lying outside the 99 °/₀ limits is of course quite insignificant. The figures in the last row of the Table show how very far out we should be in relying on the hypothesis (*c*); instead of the expected 3 cases in which the observed value of r should lie outside the 99 °/₀ limits we find 36! This difference is largely due to the total inadequacy of this method of prediction when the size of the second sample is greater than that of the first.

TABLE XII.

Frequencies of values of r outside certain ranges.

		90 per cent. Range	99 per cent. Range
Frequencies to be expected in 300 samples	Mean	30	3
	Standard Deviation	5·20	1·72
Frequencies observed with limits fixed by using	(*a*) Complete Bayes' Theorem	29	6
	(*b*) ν_1' and σ and Normal Curve	24	5
	(*c*) ${}_0\nu_1'$ and σ_0 and Normal Curve	88	36

These results seem to me to bring out remarkably well the value of Bayes' Theorem; to have found in 300 samples only 6 in which r falls outside the predicted limits when we were certainly prepared to find 3, has been seen not to be an improbable result judged on the basis of probable errors calculated on the assumption of perfect sampling. And when we are not dealing with bags of balls but have taken 300 samples from less clearly defined populations, the result is certainly satisfactory. The difference between the findings on hypotheses (*a*) and (*c*) serves to emphasise this point, and shows that it cannot be argued that *any* form of approximate rule would lead to as good results.

---

* For example, for $\begin{cases} \beta_1 = 0{\cdot}0,\ \beta_2 = 3{\cdot}0, \text{ the total range is } (1{\cdot}64 + 1{\cdot}64)\,\sigma = 3{\cdot}28\,\sigma, \\ \beta_1 = 1{\cdot}2,\ \beta_2 = 4{\cdot}6, \text{ the total range is } (1{\cdot}26 + 1{\cdot}91)\,\sigma = 3{\cdot}17\,\sigma. \end{cases}$
For the 99 °/₀ range the position is reversed.

There is however a more stringent method of analysing the results of these 300 samples. The second samples of m may be considered as 300 samples each drawn at random from a different distribution whose moments are known—i.e. from the 300 theoretical distributions of r in second samples. Suppose in general that a sample of N has been drawn, one individual taken randomly from each of N frequency distributions represented by

$$f_1(a),\ f_2(b),\ f_3(c)\ \ldots\ldots,$$

where there is no correlation between the variates $a, b, c, \ldots$. Further, suppose that for each distribution the variate is measured from the mean of that curve, and that the variate scale is such that the standard deviation is unity. Let square brackets imply a mean value for a particular distribution, and the sign S, a summation for all N distributions.

Then $$[a_s] = 0 \text{ and } [a_s^2] = 1.$$

A particular sample will contain N individuals with variate values, $a_s, b_t, c_r, \ldots$ etc. and its qth moment will be

$$M_q = \frac{1}{N}(a_s^q + b_t^q + \ldots).$$

If repeated samples of N are taken in the same manner*, then the mean value of M_q among these samples will be

$$\bar{M}_q = \frac{1}{N} S([a_s^q]) = \frac{1}{N} S({}_a\mu_q) \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xxii)},$$

while the square of the standard deviation of Mq will be

$$\begin{aligned}\Sigma_q^2 &= (\bar{M}_q^2) - (\bar{M}_q)^2 \\ &= \frac{1}{N^2}\{S[a_s^{2q}] + 2S'[a_s^q b_t^q]\} - \frac{1}{N^2}\{S({}_a\mu_q)\}^2,\end{aligned}$$

where S' indicates the sum for all possible different combinations of the variates, a and b, a and c, b and c, etc.; there will be $\frac{N(N-1)}{2}$ such combinations. As there is no correlation between a_s^q and b_t^q, it follows that

$$[a_s^q b_t^q] = [a_s^q][b_t^q] = {}_a\mu_q \times {}_b\mu_q.$$

Therefore $$\frac{2S'}{N^2}[a_s^q b_t^q] = \frac{2}{N^2} S'({}_a\mu_q \times {}_b\mu_q)$$

$$\begin{aligned}&= \frac{1}{N^2}({}_a\mu_q + {}_b\mu_q + \ldots)^2 - \frac{1}{N^2}({}_a\mu_q^2 + {}_b\mu_q^2 + \ldots) \\ &= \frac{1}{N^2}\{S({}_a\mu_q)\}^2 - \frac{1}{N^2}\{S({}_a\mu_q^2)\}.\end{aligned}$$

---

* One individual from each distribution, where the total frequency of each is supposed infinite.

Hence $$\Sigma_q^2 = \frac{1}{N^2} S({}_a\mu_{2q}) - \frac{1}{N^2}\{S({}_a\mu_q^2)\} \qquad \text{.....................(xxiii)},$$

where q may now be given the special values 1, 2, 3 and 4.

In view of the initial conditions,

$$\begin{array}{ll} {}_a\mu_1 = [a_s] = 0, & {}_a\mu_2 = [a_s^2] = 1, \\ {}_a\mu_3 = \sqrt{{}_a\beta_1}, & {}_a\mu_4 = {}_a\beta_2, \\ {}_a\mu_6 = {}_a\beta_4, & {}_a\mu_8 = {}_a\beta_6, \end{array}$$

so that we obtain from (xxii) and (xxiii),

$$\left.\begin{array}{ll} \bar{M}_1 = 0, & \Sigma_1 = \dfrac{1}{\sqrt{N}} \\ \bar{M}_2 = 1, & \Sigma_2 = \dfrac{1}{\sqrt{N}}(\text{Mean } \beta_2 - 1)^{\frac{1}{2}} \\ \bar{M}_3 = \text{Mean } \sqrt{\beta_1}, & \Sigma_3 = \dfrac{1}{\sqrt{N}}(\text{Mean } \beta_4 - \text{Mean } \beta_1)^{\frac{1}{2}} \\ \bar{M}_4 = \text{Mean } \beta_2, & \Sigma_4 = \dfrac{1}{\sqrt{N}}(\text{Mean } \beta_6 - \text{Mean } \beta_2^2)^{\frac{1}{2}} \end{array}\right\} \text{......(xxiv)},$$

where I have written "Mean" for $\frac{1}{N}S(\quad)$, implying the mean value of a frequency constant among the N distributions.

The values of β_1 and β_2 having been calculated already it was easy to obtain the values of $\bar{M}_q$ and Σ_q and to compare with these the observed values of M_1, M_2, M_3, M_4. It was first found from the original frequencies (which are *grouped* in Table XII) that

$$\text{Mean } \beta_1 = \cdot 4011, \quad \text{Mean } \beta_2 = 3\cdot 4388,$$
$$\text{Mean } \sqrt{\beta_1} = \cdot 4562, \quad \text{Mean } \beta_2^2 = 13\cdot 2418.$$

And then approximately,

$$\text{Mean } \beta_4 = 25\cdot 1, \quad \text{Mean } \beta_6 = 306,$$

which are the values of β_4 and β_6 corresponding to $\beta_1 = \cdot 4011$, $\beta_2 = 3\cdot 4388$ (or the *mean* values) in Pearson Type curves*. From these $\bar{M}_q$ and Σ_q were computed as shown in Table XIII. The observed values of M_q were calculated from the frequency distribution of Table XIV, formed by entering as variate each of the 300 values of $z = \frac{r - \nu_1'}{\sigma}$, where r is as usual the observed number of "marked" individuals in the second sample, and ν_1' and σ the mean and standard deviation of the corresponding distribution of Bayes' Theorem. M_1, M_2, M_3 and M_4 are given in Table XIII, both for the 300 observations and for 299, the latter excluding one, No. 257, in which the deviation of r from the mean was 4·8 times the standard deviation†.

---

* Interpolated from a new Table of the higher β's computed by Professor Yasukawa, which will shortly be published.

† For No. 257, p and r were the number of persons in random samples from "Who's Who," 1899, who were authors (as shown by quoting publications). A first sample gave 86 out of 200, a second only 14 out of 100. Possibly the pocket recorder may have been in error, but the observation was accepted at the time of record as presumably accurate, and cannot be altogether rejected.

TABLE XIII.

Moments of Distribution of z on Bayes' Hypothesis.

	Mean values in repeated series of samples		Observed values of M_q	
q	$\bar{M}_q$	Σ_q	For the 300 samples	For 299 samples
1	·0000	·0577	− ·0173	− ·0010
2	1·0000	·0902	1·0929	1·0163
3	·4562	·2869	− ·2388	+ ·1538
4	3·4388	·9879	4·7817	2·8700

TABLE XIV.

Frequency Distribution of $z = \frac{r - \nu_1'}{\sigma}$ *observed among the* 300 *double Samples.*

Limits of z	Frequency	Limits of z	Frequency
−5·0 to −4·8	1	+0·0 to +0·2	18
⋮ ⋮	⋮	+0·2 „ +0·4	17
−3·0 to −2·8	—	+0·4 „ +0·6	18
−2·8 „ −2·6	2	+0·6 „ +0·8	15
−2·6 „ −2·4	—	+0·8 „ +1·0	17
−2·4 „ −2·2	1	+1·0 „ +1·2	18
−2·2 „ −2·0	3	+1·2 „ +1·4	17
−2·0 „ −1·8	1	+1·4 „ +1·6	4
−1·8 „ −1·6	4	+1·6 „ +1·8	6
−1·6 „ −1·4	10	+1·8 „ +2·0	1
−1·4 „ −1·2	14	+2·0 „ +2·2	3
−1·2 „ −1·0	16	+2·2 „ +2·4	2
−1·0 „ −0·8	14	+2·4 „ +2·6	2
−0·8 „ −0·6	20	+2·6 „ +2·8	1
−0·6 „ −0·4	28	+2·8 to +3·0	1
−0·4 „ −0·2	24		
−0·2 to −0·0	22	Total ...	300

M_1 and M_2 are very close to $\bar{M}_1$ and $\bar{M}_2$, the "expected" values, if we omit the double sample No. 257, and if it is included $\bar{M}_2 - M_2$ is still only of the same order of the standard deviation, or 1½ times the probable error. With only 300 samples the values of Σ_3 and Σ_4 are so large that little useful information can be drawn from the observed values of M_3 and M_4. The inclusion of the single observation with a deviation of 4·8σ changes the positive skewness to negative, and increases M_4 from 2·87 to 4·78, but little importance can be attached to this although it suggests that the observed distribution of Table XIV is not as skew as we should have anticipated. The more skew the theoretical distribution of r, the more skew should be the distribution of z. Taking only those 47 samples in which the β_1 of the r distribution was greater than 0·6, results were obtained as shown in Table XV.

TABLE XV.

Moments of Distribution of z in Samples where $\beta_1 > 0{\cdot}6$.

	Mean values in repeated series of samples		Observed value of M_q in 47 samples
q	$\overline{M}_q$	Σ_q	
1	0·000	·146	− ·083
2	1·000	·307	·994
3	1·28	1·27*	+ ·47
4	5·42	6·16*	3·24

As far as any conclusions can be drawn from a result in which the values of Σ_q are so large, we may note that:

(1) M_1 and M_2 are again very near the expected values, showing that there is a marked consistency in the results—the variation in those particular samples in which β_1 is large (or the cases in which the p of the first sample of n was very small) is no more and no less than throughout the whole series of 300.

(2) As it should be, the value of M_3 is here greater than in the case of the whole 300 samples, but again it is considerably below the expected value. This divergence may perhaps be partly due to the fact that the distribution of M_3 is probably skew, so that the observed M_3 may lie nearer its modal value than it does to its mean.

Now let us consider what results would have been obtained had we made use of the approximate prediction rule, taking, as on p. 419,

$$_0\nu_1' = \frac{mp}{n}, \qquad \sigma_0 = \sqrt{\frac{mpq}{n^2}},$$

and supposing all the curves of distribution of r to be Normal. If this method of prediction was adequate we should be justified in comparing the moments of the observed distribution of

$$z' = \frac{r - {}_0\nu_1'}{\sigma_0}$$

with the mean values $_0\overline{M}_q$ and $_0\Sigma_q$ given in Table XVI†. It has been necessary to omit six cases in which p was zero but not r, so that z' was infinite; the crude theory does not allow for these. The last column of the Table again omits observation No. 257, in which the value of r had here a deviation of nearly $6\sigma_0$ from the expected value. A comparison of Table XIII with XVI shows how

---

* These two values lying beyond the range of the existing tables were calculated from the finite difference equations between successive β's of Pearson's curves.

† Here Mean $\beta_1 = 0$, Mean $\beta_2 = 3$, Mean $\beta_4 = 15$, Mean $\beta_6 = 105$, or the Normal curve values when standard deviation is unity.

TABLE XVI.

Moments of Distribution of z' on modified hypothesis.

	Mean values in repeated series of samples		**Observed values of ${}_0M_q$**	
q	${}_0\overline{M}_q$	${}_0\Sigma_q$	**For the 294 samples**	**For 298 samples**
1	·0000	·0583	+ ·1701	+ ·1904
2	1·0000	·0825	2·8252	2·7201
3	·0000	·2259	+ ·4658	1·1329
4	3·0000	·5714	33·4501	29·7044

hopelessly inadequate are the methods of prediction employed in the latter case; ${}_0M_3$, ${}_0M_2$ and ${}_0M_4$ have increasingly impossible values, and the value of ${}_0M_1$ (+ ·1701) is far less satisfactory than M_1 (− ·0173), which is a point of some importance. It shows that not only are better predictions obtained by using the standard deviation σ (from Equation (iv)) rather than σ_0, but that $\frac{m(p+1)}{n+2}$ is definitely a better prediction value to take for r in the second sample than is $\frac{mp}{n}$.

It has been shown that for the samples with high values of β_1 and β_2, M_2 remains close to the expected value of 1·0. I have applied two other tests of consistency. The accuracy of the expression

$$= \sqrt{\frac{(p+1)(q+1)\,m\,(n+m+2)}{(n+2)^2\,(n+3)}} \qquad \text{(from (iv))}$$

as a measure of dispersion has been checked by calculating the value of M_2 for the whole 300 samples. It is possible that heterogeneity might lie concealed under this general measure; for example, an unnecessarily large allowance for variation made when σ is large, which is usually the case when $m > n$, might be balanced in the long run by a too small allowance in other cases, or *vice versa*.

The coefficient of correlation was therefore found between σ and z; it was $-\cdot 0268 \pm \cdot 0389$, while $\eta_{z.\sigma} = \cdot 2352$ against a mean value to be expected in cases of no correlation of ${}_0\bar{\eta}_{z.\sigma} = \cdot 2082 \pm \cdot 0389$. There is therefore no significant relation, and σ appears to be a satisfactory measure of variation whether the limits of prediction are great or small. Finally I have considered the 129 cases in which the size of the second sample was greater than the first; among these there was great variety from the case where m was only a few units greater than n, to the extreme case where $m = 10n$ (e.g. $n = 20$, $m = 200$). It is found that

in 14 cases the value of r lies outside the 90 °/ₒ limits,

in 3 cases „ „ „ 99 °/ₒ limits.

These frequencies are a little greater than the expected numbers of 12·9 and 1·3, but on further analysis it is found that

$$M_1 = +\cdot 0457 \text{ compared with } \bar{M}_1 = \cdot 0000 \pm \cdot 0594^*,$$
$$M_2 = +\cdot 9934 \quad \text{,,} \quad \text{,,} \quad \bar{M}_2 = 1\cdot 0000 \pm \cdot 0927^*,$$

which is very satisfactory.

We may therefore conclude that as far as the results of this series of experimental samplings are concerned, the use of the measures of expected frequency (ν_1') and variation (σ) in second samples obtained from the simple Bayes' hypothesis in which $F(P)$ is taken as constant, is entirely justified by the frequencies in the observed second samples. It is a powerful feature of Bayes' Theorem that it allows for the first sample being much smaller than the second, and the experimental results suggest that this allowance is neither more nor less than adequate. The accuracy of the expressions for the higher moments cannot be certainly checked owing to the very high probable errors of constants determined in this special manner from only 300 samples, but the results of the sampling of Series I certainly suggest that the theoretical measures of skewness are likely in the long run to be of real use in prediction.

8. In the preceding section it has been assumed that the distribution $F(P)$ was rectangular, and the success obtained in the prediction of limits in second samples has shown that this assumption was justified. It is however interesting to consider the observed distribution of the ratios $(p+r)/(n+m)$ a little more closely; the form of distribution was shown in Fig. 3. In the first place, this has been fitted with a Type I curve of the form of Equation (v), choosing a so as to make the second moments of curve and observation agree. The curve obtained is

$$y = 22\cdot 80 x^{-\cdot 1336} (1-x)^{-\cdot 1336} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xxv)},$$

and is shown in the figure. The fit is clearly very rough, largely because of the irregular jumps in frequency for values of $(p+r)/(n+m)$ centred at ·15 and ·85. As the distribution is based on only 300 samples and this irregular frequency contains 39 cases against 32 in each of the neighbouring groups, no valid conclusions can be drawn from this result alone.

There was however a similar maximum in the distribution of $p+r$ found in Series I (see Fig. 2), and it is possible that the feature has some real significance. For example, it would be possible to advance the hypothesis that if a free and unprejudiced selection of random samples could be taken from among the populations of experience, the distribution of P would be found to be a U-curve following approximately the distribution of Fig. 3 between $P = \cdot 15$ and ·85, and then rising steeply to $P = 0$ and 1. In this event the observed shortage of cases with $(p+r)/(n+m)$ below ·15 and above ·85 would be due to my own personal equation as an observer, to some unconscious prejudice against selecting freely for the experiment, propositions with which a very low or high value of P is associated. On this hypothesis we might fit a portion of a Type I curve having limits at $P = 0$ and $P = 1$, to the observed frequencies in the range ·15 to ·85, and claim that this was the true $F(P)$ distribution*. Such a curve has been fitted by the method of least squares, using the logarithms of ordinates. It has the equation

---

* Taking the probable errors as ± ·67449 times Σ_1 and Σ_2.

$$y = 9{\cdot}838x^{-{\cdot}6620}(1-x)^{-{\cdot}6620} \quad \text{.......................(xxvi)},$$

and is shown in Fig. 3. It is certainly possible that in collecting material for this special purpose there would be a tendency not to fix on examples in which P is very low or very high, simply because they would not occur to one; while the statistician in practice dealing with material that came before him more or less by chance might find a more definite U-shape in his distribution $F(P)$. Whether this is so, could only be shown by the test of wider experience, but I doubt that the difference could be so great as between my observations and the curve of Equation (xxvi). The area under this curve is to the area of the histogram in Fig. 3 as 1025 to 300, so that if the former were the typical curve the statistician in practice must be supposed to come across 7 *additional* populations with very high or low values of P for every 3 that I have met.

But let us suppose that this were so, and that a statistician were to make use of Bayes' Theorem with $F(P)$ constant for purposes of prediction when dealing with a large series of populations in which $F(P)$ was really represented by a curve similar to (xxvi). Would his predictions be entirely invalid? It has been shown that the introduction of a U-shaped $F(P)$ curve will only seriously modify the moments of the expected distribution of r in second samples when p is but a few units. Only such cases need be considered; the values of constants, limits, etc. for three of these taken from the 300 double samples are given in Table XVII:

No. 92. As has been shown before, a negative α gives a lower mean and standard deviation and higher values of β_1 and β_2, or greater skewness†. In this particular case we note a very considerable difference in the value of these constants, according to whether $\alpha = 0$ or $-{\cdot}66$, so that the curves of distribution of r in second samples do not correspond closely. But the statistician will not be dealing with repeated samples in which $n = 60$, $m = 30$, $p = 1$, and so for most practical purposes will not wish to make use of the whole form of the curve. He will be content to obtain some limit of range within which he may expect r to fall. The later columns of Table XVII give the deviations of the limits of the central 90 % and 99 % of area taken from the Tables of the Appendix, and also these limits converted into terms of frequency of "successes" in the second sample, by multiplying by σ and adding to the Mean. In the first place, consider the lower limits; their values are here of no great importance, for they would not be made use of. The modal value of r is 0‡, and the Type I curve cannot be expected to correspond exactly with the abrupt start of the hypergeometrical histogram. In such cases therefore the lower limit of r would naturally be taken as zero. For the upper limits of r, we have

(1) 90 %: 3·3 for $\alpha = 0$, 2·7 for $\alpha = -{\cdot}66$,
(2) 99 %: 5·5 for $\alpha = 0$, 4·8 for $\alpha = -{\cdot}66$.

---

* Or rather the curve when reduced so that the total area below it is unity.

† The points representing the Bayes' hypergeometrical series lie so near to the Type III line, that the skewness is closely proportional to $\sqrt{\beta_1}$.

‡ The first two terms of the series representing $C_{p,r}$ are proportional to 1 and $m(p+1)/(q+m)$.

TABLE XVII.

Examples showing the effect of the introduction of a U-shaped $F(P)$ curve with a value of $a = -\cdot 66$.

Observation	n	m	p	a	Mean	σ	β_1	β_2	Deviations, from Tables for 90 per cent. Range		Deviations, from Tables for 99 per cent. Range		90 per cent. Limits		99 per cent. Limits		Observed r
									−	+	−	+					
Number 92	60	30	1	0	·968	1·169	2·33	5·99	1·05	2·01	1·13	3·89	(− ·3)	3·3	(− ·4)	5·5	0
				− ·66	·663	·976	3·56	7·66	—	2·04*	—	4·22*	—	2·7	—	4·8	
Number 135	20	20	1	0	1·818	1·737	1·48	4·75	1·17	1·96	1·32	3·60	(− ·2)	5·2	(− ·5)	8·1	1
				− ·66	1·296	1·508	2·38	5·98	1·04	2·02	1·10	3·90	(− ·3)	4·3	(− ·4)	7·2	
Number 157	150	150	7	0	7·895	3·842	·47	3·62	1·42	1·82	1·90	3·18	2·4	14·9	·6	20·1	6
				− ·66	7·307	3·712	·52	3·69	1·41	1·83	1·87	3·22	2·1	14·1	·4	19·3	

* Values obtained by extrapolation from Tables.

These are certainly different, but those for $\alpha = -\cdot 66$ are *less* than for $\alpha = 0$; that is to say, if we use the Bayes' hypothesis assuming that $\alpha = 0$ when we are really sampling from populations among which P follows a distribution in which $\alpha = -\cdot 66$, the upper limits to r that we predict will be *higher* than their real values, and we shall therefore be erring on the side of safety.

No. 135. In this case the curves are not quite as skew as before; but the first two terms in the hypergeometrical series are 1 and 40/39, so that the Type I curve is still not adequate at this tail, and the lower limit of r is again 0. At the other tail, the upper limit is in both cases about a unit higher when $\alpha = 0$ than when $\alpha = -\cdot 66$; that is to say, we should again be on the cautious side in using the simple Bayes' Theorem for prediction.

No. 157. Here p is larger $(=7)$ and, except for the Means, the two sets of moments have almost reached the point where the difference is quite insignificant. The histogram is no longer J-shaped*, the Type I curve will give a better fit at the commencing tail and the lower limit for the 90 % range is not at zero. Here the value for $\alpha = 0$ (2·4) lies within that for $\alpha = -\cdot 66$ (2·1), but the difference is not serious, and in practice, only integral values of r being possible, we should take 2 as the lower limit in both cases. Similarly for the 99 % range, we should take 0 as the lower limit. For the upper limits, we find again that the curve for $\alpha = 0$ gives us a unit more on the safe side.

The position is therefore as follows. If we form our limits of prediction on the assumption that $\alpha = 0$, then even in the extreme case when $\alpha = -\cdot 66$, there will only be a sensible error when p is small. When this is so the upper limit of prediction for r (whether for 90 % or 99 % range) will be somewhat higher than it need be, and while the lower limit will also be a little higher†, with small values of p this will generally be taken as zero. The position is represented diagrammatically on an exaggerated scale in Fig. 4; the curve for $\alpha = 0$ has a higher mean, a larger standard deviation, but less skewness than that for a negative α.

The position would not be quite so satisfactory if α were a positive fraction and the curve $F(P)$ as (*b*) of Fig. 1. Here the changes in the moments would be of opposite sign, since to the first order they depend on $\frac{\alpha}{p+1}$; that is to say, in assuming that $\alpha = 0$, we should base our predictions on a curve having too low a mean, too small a standard deviation and too great skewness. The predicted upper limit for r (for the 90 and 99 % areas) would now be a little too *low*. But at present there is no evidence that this experience is likely to be met with, for it seems probable that the $F(P)$ curve, if not horizontal, will be concave rather than convex.

The particular form of distribution of $F(P)$—a Type I curve—has been taken for purposes of illustration; it should not be supposed for a moment that the distributions of P among the populations of other observers' experience will be found *a posteriori* to follow such a curve at all closely. My own series of observations have not in fact done so. But the discussion of this illustration in some detail serves two purposes; it shows how a possible form of divergence from linearity in the distribution of P will affect the moments of the distributions of r, and from the comparatively unimportant nature of these modifications gives us confidence to apply Bayes' Theorem to predict the limits of r in future sampling.

---

* The first three terms are approximately as 1, 4, 8.

† But not as much higher as for the upper limit.

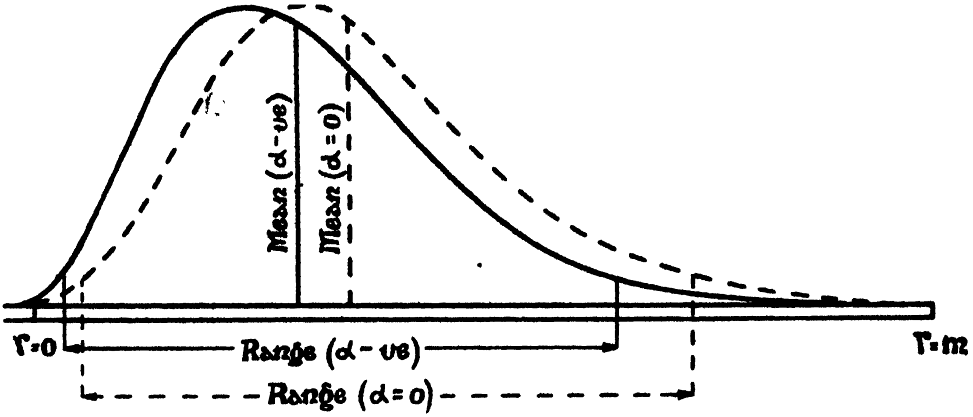


Fig. 4. Diagram representing the limits of percentage frequencies in curves for predicting the values of r in second samples.

It is now possible to consider more clearly the problem that so much exercised the 19th century writers, that is to say, the case where p (or q) is 0. In the particular case when $F(P)$ was taken as a Type I U-curve it was necessary to make the limitation that there should be no populations with $P=0$ or 1 exactly, in order to reach the integrals of (i) from the summations of (i a). But for the case $F(P)=$ constant this was not necessary, and in general there is no reason why the populations sampled should not contain a finite proportion in which $P=0$ or 1. In both the present series of experiments there were very few cases in which this could be considered to be so*, for one naturally avoided absolutely impossible or absolutely certain propositions, realising that but little mental effort in selecting propositions would be needed to swamp the whole of the experiment with cases of absolute certainty or the reverse. On the other hand in practical sampling cases will occur when $P=0$ or 1 although the observer can neither tell that this is so *a priori* or *a posteriori*. For example, in medical work it may be desired to test the result of a certain treatment; the observer may have good grounds for believing that the conditions in two samples are the same, but yet be quite uncertain *a priori* whether the treatment is going to have any effect or not. On finding no cures in the first sample he will still be uncertain, but Bayes'

---

* There was one case I know when this was so, in which I searched in samples from the list of members of the Hackney Stud Society for the number of lady members, to find afterwards that they were included in a separate list. But even here there was a chance of a secretarial error including one among the men.

Theorem will enable him to predict limits of range within which the number of cases is likely to fall in the second sample. All that is essential in the long run for accuracy of prediction through a series of varied tests is that among the population sampled in these tests the distribution of P will not differ more from linearity than in the cases that have been considered. The fact that in my experiments there were very few cases with $P=0$ or 1 exactly is of little consequence in practice, for there is no sensible difference in prediction between the cases where $P=0$ and $P=\cdot 001$. In the first series of experiments, it was found that the observed distribution of r in second samples where p had been 0 in the first sample, gave a very good fit to theory*. Further, it must be remembered that if $P=0$, both p and r will be 0, so that r will certainly fall *within* the predicted limits. In fact if some of the early writers had found time to divert themselves by examining the values of r in samples from a whole series of populations in which an event had never been known to occur before, it is true that the Bayes' Theorem distribution of r would not have been borne out, but had they taken reasonable caution to ensure the stability of their populations, I think they would have found in the long run that in at least 99 °/₀ of cases r would lie between the 99 °/₀ limits of frequency obtained as suggested in this paper. As for the cases in which $n=1$ and $m=1$, these hardly belong to practical statistics. The Type I curves for distribution of r would of course be inappropriate†, yet in so far as the sampling was carried out from populations among which $F(P)$ varied within the limits we have considered, I see no reason to doubt that in the long run the predictions of Bayes' Theorem would be carried out.

9. Finally let us consider as an illustration the 20 cases selected randomly from the 300 double samples of Series II, which are described in Table XVIII. Take the first as an example: in the Euston Road, on a certain day, out of 200 men passed, 26 have been observed to carry umbrellas; within what limits shall we expect to find the number falling in the next 150 that are observed? From n, m and p we calculate the moments of the distribution of r from Equations (iv); the linear relation between β_1 and β_2 of (xix) then forms a useful check on the computation. Next we enter the Tables of the Appendix with β_1 and β_2 and find the deviations from the mean,

(*a*) for the central 90 °/₀ of area (Tables (1) and (2)) $-1\cdot 55$, $+1\cdot 74$,

(*b*) „ „ 99 °/₀ „ (Tables (3) and (4)) $-2\cdot 24$, $+2\cdot 87$.

Then the corresponding limits of range are

$$(a)\quad \nu_1' - 1\cdot 55\sigma = 11\cdot 5^+, \quad \nu_1' + 1\cdot 74\sigma = 29\cdot 6,$$

$$(b)\quad \nu_1' - 2\cdot 24\sigma = 7\cdot 8, \quad \nu_1' + 2\cdot 87\sigma = 35\cdot 8.$$

That is to say in terms of probability the chances are at least 9 to 1 that r will lie between 11 and 30, and at least 99 to 1 that it will lie between 7 and 36. Actually the observed value was 13.

---

[*] See Tables IV and VI.

[†] For n and m between 1 and 15 the terms of the hypergeometrical series can be readily computed or use may be made of Table XLVIII in *Tables for Statisticians and Biometricians*.

If a more rapid but approximate method of prediction is required, we can calculate only ν_1', σ and β_1, and assuming the curve of distribution of r to be adequately represented by Type III[*], take as the deviations from the Appendix Tables the bottom figures in the appropriate β_1 column. Or more roughly still we may compute only ν_1' and σ, and assuming the curve of r to be Normal, take the limits of range at

$$\nu_1' \pm 1{\cdot}6449\sigma = 11{\cdot}0^+ \text{ and } 29{\cdot}1\,; \text{ and } \nu_1' \pm 2{\cdot}5758\sigma = 5{\cdot}9 \text{ and } 34{\cdot}2.$$

It was seen that this last method gave satisfactory results for the 300 double samples[†]; but we cannot be certain that it would invariably do so and the wiser course is to make use of the full rule which at any rate on theoretical grounds is definitely more accurate in the long run.

The full rule has therefore been employed in calculating the 90 °/ₒ limits given in Table XVIII. It will be seen that in no case does r fall outside these, although we should have been prepared to find this occur in 2 cases (or 10 °/ₒ). There is no need to discuss the results in detail, but it should be specially noted how, in those cases where the first sample is smaller than the second, Bayes' Theorem provides a very large standard deviation, so allowing for the greater uncertainty of prediction. In the two cases where the lower limit is negative and given in brackets, it would be naturally taken as at 0. As for the complete 300 cases, a measure of the agreement of observation with theory may be obtained from the moments of the distribution of $z = (r - \nu_1')/\sigma$. The values of z are given in the last column of the Table. Using the previous notation[‡],

$$M_1 = +{\cdot}0289, \qquad M_2 = {\cdot}5652,$$

while $$\bar{M}_1 = {\cdot}0000 \pm {\cdot}1508, \quad \bar{M}_2 = 1{\cdot}0000 \pm {\cdot}2355.$$

Thus the difference between the observed and "expected" values of r is in the long run neither too great nor too small, while the actual variation about the expected values is exceptionally low, although I think no significance can be attached to this. The method of random selection employed in nearly all the examples of Table XVIII has been by alphabetical arrangement. This may appear a form of selection not so very different from drawing balls out of a bag, and it might perhaps be thought that so much stability will not be found among the populations which we are sampling in practical work. To test this point I have analysed the double samples Nos. 1—52, which were all collected in the street, or in trains, buses, etc., that is to say without any form of mechanical selection[§]. In these 52 cases, r lies 3 times outside the limits of the central

---

[*] For my particular series of observations the values of β_2 given by the regression line (xxi) would be more accurate in the long run.

[†] p. 419.

[‡] pp. 420—421, taking the probable errors of $\bar{M}_1$ and $\bar{M}_2$ as $\pm{\cdot}67449\,\Sigma_1$ and $\pm{\cdot}67449\,\Sigma_2$, and obtaining Σ_2 by using for Mean β_2 the mean value for all 300 cases, i.e. 3·4388.

[§] Here are some examples of characters observed: No. 1, Bus drivers with moustaches; No. 5, Men wearing glasses in the street; No. 18, Morris cars among all private cars; No. 36, Men in the Tube reading newspapers; No. 48, Men without collars met in passing along a certain street; No. 50, Copper coins with King Edward's head on.

TABLE XVIII.

Example of 20 cases from among the double samples of Series II.

No.	Population	Sub-class (p and r)	n	m	p	Mean	σ	β_1	β_2	Limits of Range of central 90 per cent.		r	z
										Lower	Upper		
15	Men met in the street (within half an hour)	Carrying umbrellas	200	150	26	20·050	5·488	·108	3·115	11·5^{+}	29·6	13	−1·285
30	Men met in the street	Wearing bow ties	185	62	7	2·652	1·834	·673	3·827	0·2	6·1	2	− ·356
45	Couples met in the street	Consisting of 2 men	53	29	19	10·545	3·173	·029	2·892	5·5^{+}	16·0^{-}	11	+ ·143
60	Names in *Telephone Directory* ...	Business or occupation quoted ...	500	100	230	46·016	5·453	·000	2·972	37·1	55·0^{-}	46	− ·003
75	Streets, etc. in Postal District List ...	District numeral a "1"; e.g. S.E. 1 ...	100	20	23	4·706	2·065	·123	3·011	1·5^{+}	8·3	2	−1·311
90	Top lines of pages in *Rewards and Fairies*	Containing 2 verbs only	40	70	14	25·000	6·470	·033	2·904	14·8	36·1	28	+ ·464
105	Individuals in *Dictionary of National Biography*	Died in 17th century	40	30	5	4·286	2·480	·446	3·445	0·8	8·8	4	− ·115
120	*U.S.A. Census*; towns, villages, etc. in Pennsylvania	Ending with letter "a"	170	100	10	6·395	3·068	·369	3·464	1·9	11·9	8	+ ·523
135	*U.S.A. Census*; towns, villages, etc. in California	With population under 300	20	20	1	1·818	1·737	1·480	4·746	(− ·2)	5·2	1	− ·471
150	Mares in *General Thoroughbred Stud Book*	With grey coats	30	300	0	9·375	9·559	3·324	7·667	0·0^{+}*	28·6*	2	+ ·772
165	Mares in 1913 Volume of *Hackney Stud Book*	Foaled in 1908 or earlier	50	200	10	42·308	12·594	·147	3·104	23·0	64·3	53	+ ·849
180	Owners of cattle in *Guernsey Herd Book*	Titled	28	28	3	3·733	2·460	·642	3·676	0·4	8·3	5	+ ·515
195	R.N. Officers in *Navy List*, 1905 ...	In the Marines	300	60	19	3·973	2·105	·327	3·358	0·9	7·8	6	+ ·963
210	Individuals in French *Biographie Universelle*	Born in 19th century	80	30	2	1·098	1·195	1·720	5·142	(− ·3)	3·5^{-}	0	− ·919
225	Peerages in Burke's *Peerage*	Created in 19th century	70	40	31	17·778	3·893	·003	2·893	11·4	24·2	22	+1·085
240	Individuals in Burke's *Landed Gentry* 1875	Unmarried	100	40	25	10·196	3·236	·070	3·001	5·1	15·8	6	−1·297
255	Individuals in *Who's Who*, 1899 ...	No Club given	100	200	40	80·392	11·873	·006	2·949	61·0^{+}	100·1	82	+ ·135
270	Individuals in *Clergy Directory*, 1905	With Cambridge degree	500	50	156	15·637	3·434	·017	2·971	10·1	21·4	15	− ·186
285	*Commercial Telephone Directory*; Costumiers	Fancy names	100	50	14	7·353	3·042	·203	3·192	2·8	12·7	10	+ ·870
300	Names of places in Index to *Census* of England and Wales, 1921	Localities, etc. "without legally defined boundaries"	100	50	26	13·235	3·790	·058	2·995	7·3	19·7	14	+ ·202

* Deviations obtained by extrapolation from Tables.

90 °/₀ area, when the average in repeated series of 52 would be 5·2 times; and never outside the limits of the 99 °/₀ area. For the z distribution

$$M_1 = +\cdot 0024, \qquad M_2 = \cdot 9754,$$
$$\bar{M}_1 = \cdot 0000 \pm \cdot 0935, \quad \bar{M}_2 = 1\cdot 0000 \pm \cdot 1461,$$

or M_1 and M_2 are extraordinarily close to the "expected" values, and there is no suggestion of instability.

The distribution of values of $(p+r)/(n+m)$ among the 20 double samples of Table XVIII is as follows: between 0·0 and 0·1, 6; between 0·1 and 0·2, 4; between 0·2 and 0·3, 4; between 0·3 and 0·4, 3; between 0·4 and 0·5, 3. This suggests a slightly U-shaped distribution of $F(P)$.

10. What then is the final position? Looking at the problem from the point of view of frequency rather than probability, I have tried to show the position which I believe Bayes' Theorem should take in the field of practical sampling. I do not start from the point of view of an *a priori* distribution of probabilities, nor do I suppose that the statistician in sampling a population is entirely ignorant of its contents. The exact form of the distribution, $F(P)$, seems to be something which each statistician can only determine for himself *a posteriori* by an examination of his own statistical experience, and Bayes' Theorem can only be accepted as providing a valuable working rule for prediction, on the assumption that among the problems with which most statisticians are confronted there is in fact a distribution of values of P between 0 and 1 whose difference from linearity is of the same order as that observed in the experiments that have been discussed in this paper. This assumption does not appear to be unreasonable, but no final conclusions can be reached until evidence has been collected from a wider source*. Again I would not suggest that the theorem should be applied to problems where we are in complete ignorance of everything but the values of n, p and m. Sufficient must be known of the factors influencing the character of the population to provide good grounds for supposing that the conditions have not changed between the drawing of the first and second sample. Certain risks must be taken or no predictions could ever be made; but the acceptance of these risks, which fall outside the scope of reasoning of the mathematical theorist, is justified by past experience in the eyes of the practical worker.

Again, the observer may sometimes have relevant information enabling him to alter or narrow his prediction limits for r. He may know for some reason that the p in his first sample is particularly far from the modal value which would be observed in repeated samples. Just as some Superman who followed the shakings of his bag of balls might tell by the aid of dynamics, of elasticity and of geometry the exact constitution of a particular sample of balls, so in our problems we may sometimes have information which tells us which, out of the universe of possible samples, our particular sample may be. But this does not detract from the value of Bayes' Theorem as providing an accurate mass result when applied to a great range of problems, and how often too is that supposed relevant information unreliable or in a shape incapable of being put into numerical form! Take for instance the 20 examples of Table XVIII; would it have been possible to fix by any other rule, limits of no wider range within which nine-tenths of the values of r would have been found to fall? It is true that these particular 20 questions have a somewhat academic ring which can hardly be avoided in a large scale *ad hoc* experiment, yet surely in our everyday work we do often have to sample populations in which there is as much reason to expect stability as in the successive pages of a book or at the two ends of a street? On this point the reader must form his own final judgment, but if he allows that the experiments are typical and that the conclusions that have been drawn are correct, then he must agree that the results have a wider bearing on the whole question of probable errors as well as on the particular problem of Bayes' Theorem. For they provide further evidence to confirm our belief in the stability of statistical ratios in the world of experience upon which so much practical work depends.

---

* One feels inclined to urge statisticians to apply Bayes' Theorem wherever possible in their work, and to record and compare their results.

In the special problem, as throughout the whole field of statistics, it is perhaps necessary to allow a slightly greater latitude in variation than the theory of probable errors based on the "bag of balls" method of sampling permits. But this allowance is not great; at worst, as in the first series of experiments, we find a general Goodness of Fit result of $P(\chi^2) = \cdot 05$ where we should have liked to find $\cdot 50$; while at best we have the values of M_1 and M_2 in the 52 cases of street sampling—$\cdot 0024$ and $\cdot 9754$ against theoretical values of $\cdot 0000 \pm \cdot 0935$ and $1 \cdot 0000 \pm \cdot 1461$.

11. *The inverse of Bayes' Theorem.* There is a problem which is often met with where the position is the inverse of that which we have considered. After finding a frequency of occurrence, p, of a certain character in a sample of n, a frequency, r, of the same character is observed in a second sample of m; what measure can be given to the probability that the two samples come from the same population? Suppose that the test is applied on N occasions to samples drawn from populations among which the distribution of P is such that the simple Bayes' Theorem can be justifiably applied. From the observed values of n, m and p we can compute in each case a limiting range outside which only k% of the values of r should lie on the Bayes' hypothesis if both samples were drawn at random from the same population. Suppose out of the N cases,

(1) in A, the two samples have been drawn from the same population, and that in a of these r lies outside the k% limits;

(2) in B, the two samples have been drawn from different populations, and that in b of these the value of r lies outside the limits. B will not equal b, for some good fits will occur when sampling from different populations.

Now if the test is applied indiscriminately to all pairs of samples observed, then when A is large $100a/A \rightarrow k$. If k be so small that we decide to take the risk of rejecting all those double samples in which the value of r lies outside the k%

limits, as almost certainly coming from different populations, then we shall be wrong in this assumption in a proportion of $a/(A+B)$ or a/N of the total cases, which is less than $k\,\%$. This is a definite criterion giving an outside limit of error. On the other hand in accepting double samples where r lies within the limits, as probably all having both samples from a single population, we shall be wrong in a proportion $(B-b)/N$ of the total cases examined. This may be a very small quantity, but we have no means of measuring the relations of B, b and N to one another, for they depend to some extent on what is meant by the term "same population," and will vary according to the proportions of A and B among our tests.

It is important also to be clear regarding the limit of error risked by rejecting cases where the observed value of r is outside the limits of the central $k\,\%$ frequency of the r-distribution curve. In terms of observed frequency the rule will only be found accurate in the long run, if the test is applied to all double samples whether the correspondence of p and r is bad or good. For suppose the test were only to be applied where a bad fit was evident from inspection, then the proportion $a/(A+B)$ would really be nearer $a/(a+b)$, which is a quantity to which we can give no measure, and is likely to be far larger than $k\,\%$. On applying the test, the observer may find it extremely unlikely that a certain frequency r should be found in a sample from the same population that has previously given him a frequency p; yet supposing he has been studying double samples from very many varied populations but has not applied the test until this occasion when he finds an obviously discordant pair, then this improbability vanishes. The position is perhaps obvious, but I think it is sometimes overlooked. If, to take a different example, in our statistical experience, we only apply the χ^2 Test for Goodness of Fit to cases where from inspection the fit looks unsatisfactory, then clearly we must be very careful what deductions we draw. For we might obtain in nearly $100\,\%$ of tests, fits with $P(\chi^2)$ less than $0{\cdot}1$, where yet in every case the sample had been drawn from the supposed population.

It is of course the old difficulty; once or twice in a thousand times an event will occur against which the odds are 999 to 1, but when it does occur we are inclined to think that something is wrong, focusing our attention on the single event and forgetting the 999 times when we have observed that it did not happen. And so in practical statistics it seems to me that measures of probability can only be usefully interpreted by a reference to statistical frequency, whether recorded in the past or to be expected in the future. And it is from this point of view that I have tried to examine Bayes' Theorem throughout the paper.

Appendix, containing Tables of Deviates of Type I Curves.

The general Type I curve

$$y = y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2} \quad\text{..........................(i)},$$

can be transformed by a change of origin and scale to the form

$$y = y_0'\,x^{s-1}(1-x)^{t-1} \quad\text{..................................(ii)}.$$

If we write

$$B_x(s, t) = \int_0^x x^{s-1}(1-x)^{t-1}\,dx \Big/ \int_0^1 x^{s-1}(1-x)^{t-1}\,dx \text{ ....(iii)},$$

then $B_x(s, t)$ is the proportion of the total frequency under the curve lying between the tail at which the origin lies and the ordinate at a distance x therefrom, or the Incomplete Beta-Function.

The four Tables appended give for curves with different values of β_1 and β_2 the deviations from the mean to the ordinates for which (1) $B_x(s, t) = {\cdot}05$, (2) ·95, (3) ·005, (4) ·995, the unit of measurement being the standard deviation of the curve. Thus when we know the mean, the standard deviation and the moment coefficients β_1 and β_2 of a particular Type I curve falling within the range of the tables, Tables (1) and (2) give the limits of the "central" 90 °/₀ of frequency, and Tables (3) and (4) of the "central" 99 °/₀. The curves are supposed drawn with positive skewness so that the deviations of (1) and (3) are really negative, those of (2) and (4) positive, the smaller deviations lying to the steep side of the curve.

The field covered by the tables is shown in the β_1, β_2 diagram of Fig. 5; it was chosen to cover the range of Type I curves likely to be met with in the application of Bayes' Theorem, but pending the construction of tables of the Incomplete Beta-Function, the tables should be of use in other problems.

For example, for the distribution represented by the expansion of the binomial $(p+q)^n$,

$$\beta_1 = \frac{1-4pq}{npq}, \quad \beta_2 = 3 + \frac{1-6pq}{npq},$$

so that for a given value of n, eliminating p and q,

$$\beta_2 - \beta_1 - 3 + \frac{2}{n} = 0 \text{ ....(iv)}.$$

That is to say, for a given n the points in the β_1, β_2 field representing the binomial for varying p's will lie on a fixed straight line parallel to the Poisson line $\beta_2 - \beta_1 - 3 = 0$ and cutting the β_1 axis at the point $3 - \frac{2}{n}$. The Poisson line, which is shown in Fig. 5, divides the field of positive and negative binomials. If n be not too small, we may take the Type I curve with the same values of β_1 and β_2 as providing a good representation of the binomial, and the four Tables will therefore give the limits within which the central 90 °/₀ and 99 °/₀ of frequency lie for a large range of positive and negative binomials.

Again, there are many problems in which the distribution in samples, of a mean or other frequency constant, is represented, not by the Normal Curve, but by a skew curve whose moments are known theoretically; within the range covered, the tables will here provide a criterion corresponding to the "probable error," enabling a judgment to be formed on the probability of the occurrence of the observed constants in a random sample from the assumed population. It should be noted that the limits of the central 90 °/₀ of frequency correspond in the case

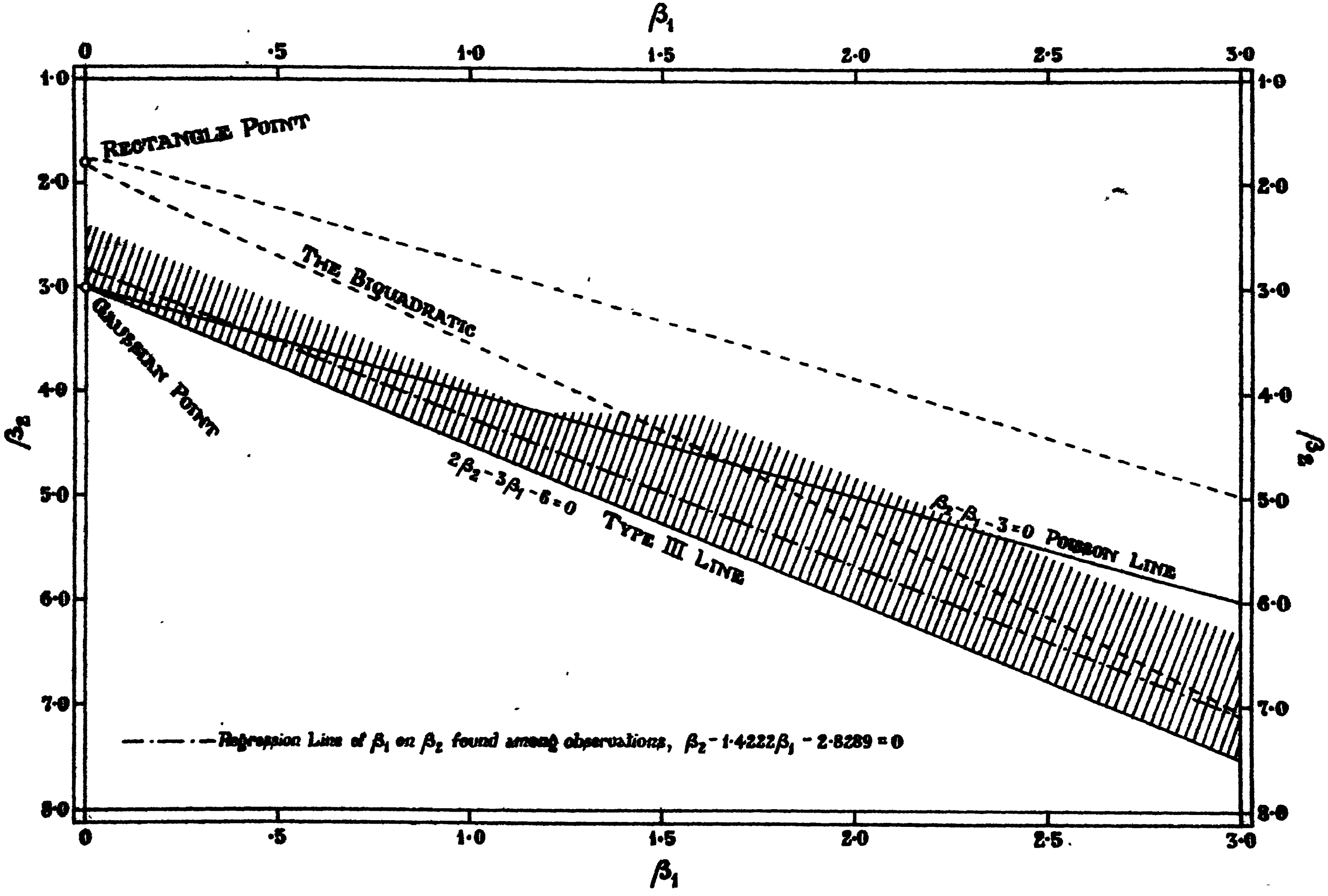


Fig. 5. Diagram of β_1 and β_2 field. The shaded area is that covered by the accompanying tables.

of the Normal Curve to a deviation of ± 2·44, and of the central 99 °/₀ to ± 3·82 times the probable error.

The tables were computed as follows:

For values of β_1 between 0·0 and 1·4 a frame of points was taken along the parallel lines

$$2\beta_2 - 3\beta_1 - 4{\cdot}8 = 0 \qquad \text{(v)},$$
$$2\beta_2 - 3\beta_1 - 5{\cdot}4 = 0 \qquad \text{(vi)},$$
$$2\beta_2 - 3\beta_1 - 6{\cdot}0 = 0 \qquad \text{(vii)}.$$

For values of β_1 between 1·4 and 3·0 the lines used were

$$2\beta_2 - 3\beta_1 - 3{\cdot}6 = 0 \qquad \text{(viii)},$$
$$2\beta_2 - 3\beta_1 - 4{\cdot}8 = 0 \qquad \text{(v) } bis,$$
$$2\beta_2 - 3\beta_1 - 6{\cdot}0 = 0 \qquad \text{(vii) } bis.$$

The computation of these frame values along the lines (v), (vi) and (viii) was carried out either by quadrature, or in certain cases where the tail of the curve rose very steeply, by the expansion of (ii) in powers of x. In applying quadrature, the areas from the tail to four equidistant ordinates in the neighbourhood of the desired limit (bounding the ·05 or ·005 tail section) were calculated, and the exact position of the limiting ordinate found by backward interpolation. Equation (vii) is that of the Type III line, and in the great majority of cases the frame values along this line could be found from the Table of the Incomplete Gamma-Function.

About 35 curves were used in all. The frame values of the deviations were calculated to three places of decimals, but as the differences were in some cases still large, it was not possible, without greatly extending the work by adding further lines to the frame, to calculate the interpolated values to this degree of accuracy throughout. The tables therefore go only to two decimal places, and in a certain number of cases a unit error may occur in the last place, but this is not of serious consequence for the purpose for which they are intended.

The very slow variation, as β_1 and β_2 are changed, in the sums of the two deviations, or the range of the central 90 °/₀ and 99 °/₀ frequency areas, provided a useful check on the computation. For example, along the Type III line this range for the 90 °/₀ area decreases only from $3{\cdot}29\sigma$ to $3{\cdot}02\sigma$ as β_1 changes from 0·0 to 3·0, while for the 99 °/₀ area the range increases from 5·15 to 5·23*. In the neighbourhood of this line, (vii), the skewness increases approximately as $\sqrt{\beta_1}$, and the tables show clearly how the deviations alter rapidly for low values of β_1, but far more slowly later on. Beyond the range of the tables the deviations alter very slowly, and it was found that an extension of Tables (1) and (2) would give the deviations for the 90 °/₀ area

at $\beta_1 = 4{\cdot}0$, $\beta_2 = 7{\cdot}8$ as $-{\cdot}81$ and $+2{\cdot}11$,

at $\beta_1 = 4{\cdot}0$, $\beta_2 = 9{\cdot}0$ (the Exponential Point) as $-{\cdot}95^-$ and $+2{\cdot}00$.

---

[*] It first falls very slightly to a minimum value at about $\beta_1 = 0{\cdot}4$.

TABLE (1).

Table of Deviates of Type I Curves Measured from the Mean in Terms of the Standard Deviation. $I_x(p, q) = 0·05$.

β_1.

β_2.	0·00	0·01	0·03	0·05	0·10	0·15	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·90	1·00	1·20	1·40	1·60	1·80	2·00	2·20	2·40	2·60	2·80	3·00
2·4	1·65$^-$	1·61	1·58	1·55$^+$	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·6	1·65$^-$	1·61	1·58	1·57	1·52	1·49	1·45$^+$	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·8	1·65$^-$	1·62	1·59	1·57	1·54	1·51	1·48	1·42	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·0	1·64	1·62	1·59	1·58	1·55$^-$	1·52	1·49	1·44	1·39	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·2	—	—	—	—	1·55$^+$	1·53	1·50$^+$	1·46	1·42	1·37	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·4	—	—	—	—	—	—	—	1·47	1·43	1·39	1·35$^-$	1·30	—	—	—	—	—	—	—	—	—	—	—	—	—
3·6	—	—	—	—	—	—	—	—	1·45$^-$	1·41	1·37	1·33	1·28	—	—	—	—	—	—	—	—	—	—	—	—
3·8	—	—	—	—	—	—	—	—	—	1·42	1·39	1·35$^+$	1·31	1·27	—	—	—	—	—	—	—	—	—	—	—
4·0	—	—	—	—	—	—	—	—	—	—	1·40	1·37	1·34	1·30	1·26	—	—	—	—	—	—	—	—	—	—
4·2	—	—	—	—	—	—	—	—	—	—	—	—	1·36	1·32	1·28	1·20	1·11	—	—	—	—	—	—	—	—
4·4	—	—	—	—	—	—	—	—	—	—	—	—	—	1·34	1·31	1·24	1·15	1·06	—	—	—	—	—	—	—
4·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·32	1·26	1·19	1·10	1·01	—	—	—	—	—	—
4·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·28	1·21	1·14	1·05$^+$	·97	—	—	—	—	—
5·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·24	1·17	1·09	1·01	—	—	—	—	—
5·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·26	1·20	1·13	1·05$^-$	·97	—	—	—	—
5·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·22	1·16	1·08	1·01	·93	—	—	—
5·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·18	1·11	1·04	·97	·89	—	—
5·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·20	1·14	1·07	1·00	·93	—	—
6·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·16	1·10	1·03	·96	·89	—
6·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·13	1·06	1·00	·93	·86
6·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·15$^-$	1·09	1·03	·96	·89
6·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·11	1·05$^+$	·99	·93
6·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·08	1·02	·96
7·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·10	1·04	·98
7·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·07	1·01
7·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·03
7·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·05$^+$

Additional values of deviates: (1) $\beta_1 = 4·0$, $\beta_2 = 7·8$, ·81.
(2) $\beta_1 = 4·0$, $\beta_2 = 9·0$, ·95$^-$.

TABLE (2).

Table of Deviates of Type I Curves Measured from the Mean in Terms of the Standard Deviation. $I_x(p, q) = 0·95$.

β_1.

β_2.	0·00	0·01	0·03	0·05	0·10	0·15	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·90	1·00	1·20	1·40	1·60	1·80	2·00	2·20	2·40	2·60	2·80	3·00
2·4	1·65−	1·69	1·71	1·73	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·6	1·65−	1·68	1·71	1·73	1·76	1·79	1·81	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·8	1·65−	1·68	1·70	1·72	1·75−	1·77	1·80	1·84	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·0	1·64	1·67	1·69	1·71	1·74	1·76	1·78	1·82	1·86	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·2	—	—	—	—	1·73	1·75−	1·77	1·80	1·84	1·87	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·4	—	—	—	—	—	—	—	1·79	1·82	1·85+	1·88	1·92	—	—	—	—	—	—	—	—	—	—	—	—	—
3·6	—	—	—	—	—	—	—	—	1·80	1·83	1·86	1·89	1·92	—	—	—	—	—	—	—	—	—	—	—	—
3·8	—	—	—	—	—	—	—	—	—	1·82	1·84	1·87	1·90	1·93	—	—	—	—	—	—	—	—	—	—	—
4·0	—	—	—	—	—	—	—	—	—	—	1·83	1·85+	1·88	1·90	1·93	—	—	—	—	—	—	—	—	—	—
4·2	—	—	—	—	—	—	—	—	—	—	—	—	1·86	1·88	1·91	1·96	2·02	—	—	—	—	—	—	—	—
4·4	—	—	—	—	—	—	—	—	—	—	—	—	—	1·86	1·89	1·93	1·99	2·04	—	—	—	—	—	—	—
4·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·87	1·91	1·96	2·01	2·07	—	—	—	—	—	—
4·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·89	1·94	1·98	2·04	2·09	—	—	—	—	—
5·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·92	1·96	2·01	2·06	—	—	—	—	—
5·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·90	1·94	1·98	2·03	2·08	—	—	—	—
5·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·92	1·96	2·00	2·05+	2·10	—	—	—
5·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·94	1·98	2·02	2·07	2·12	—	—
5·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·92	1·96	2·00	2·04	2·09	—	—
6·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·94	1·98	2·02	2·06	2·11	—
6·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·96	1·99	2·04	2·08	2·13
6·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·94	1·97	2·01	2·05+	2·10
6·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·96	1·99	2·03	2·07
6·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·97	2·01	2·04
7·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·96	1·99	2·02
7·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·97	2·00
7·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·98
7·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·97

Additional values of deviates: (1) for $\beta_1 = 4·0$, $\beta_2 = 7·8$, 2·11.
(2) $\beta_1 = 4·0$, $\beta_2 = 9·0$, 2·00.

TABLE (3).

Table of Deviates of Type I Curves Measured from the Mean in Terms of the Standard Deviation. $I_x(p, q) = 0{\cdot}005$.

β_1.

β_2.	0·00	0·01	0·03	0·05	0·10	0·15	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·90	1·00	1·20	1·40	1·60	1·80	2·00	2·20	2·40	2·60	2·80	3·00
2·4	2·26	2·14	2·05	1·98*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·6	2·38	2·27	2·18	2·12	1·98	1·87	1·76*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·8	2·49	2·38	2·30	2·23	2·10	1·99	1·89	1·71	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·0	2·58	2·48	2·39	2·33	2·21	2·11	2·01	1·84	1·68	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·2	—	—	—	—	2·30	2·20	2·11	1·95^{-}	1·79	1·65^{-}	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·4	—	—	—	—	—	—	—	2·04	1·90	1·76	1·62	1·50^{-}	—	—	—	—	—	—	—	—	—	—	—	—	—
3·6	—	—	—	—	—	—	—	—	1·99	1·85^{+}	1·72	1·60	1·48	—	—	—	—	—	—	—	—	—	—	—	—
3·8	—	—	—	—	—	—	—	—	—	1·94	1·82	1·70	1·58	1·47	—	—	—	—	—	—	—	—	—	—	—
4·0	—	—	—	—	—	—	—	—	—	—	1·90	1·78	1·67	1·56	1·45^{+}	—	—	—	—	—	—	—	—	—	—
4·2	—	—	—	—	—	—	—	—	—	—	—	—	1·75^{+}	1·65^{-}	1·54	1·35^{-}	1·18	1·04	—	—	—	—	—	—	—
4·4	—	—	—	—	—	—	—	—	—	—	—	—	—	1·73	1·62	1·43	1·25^{+}	1·11	—	—	—	—	—	—	—
4·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·70	1·51	1·33	1·18	1·04	—	—	—	—	—	—
4·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·58	1·41	1·25^{-}	1·11	0·99	—	—	—	—	—
5·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·48	1·31	1·17	1·04	—	—	—	—	—
5·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·55^{-}	1·38	1·23	1·10	0·99	—	—	—	—
5·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·45^{-}	1·30	1·16	1·04	0·94	—	—	—
5·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·36	1·22	1·10	0·99	0·90	—	—
5·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·42	1·28	1·15^{+}	1·04	0·95^{-}	—	—
6·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·34	1·21	1·09	0·99	0·90	—
6·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·26	1·15^{-}	1·04	0·95^{-}	0·87
6·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·32	1·20	1·09	0·99	0·91
6·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·25^{-}	1·14	1·04	0·95^{-}
6·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·18	1·08	0·99
7·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·23	1·13	1·03
7·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·17	1·07
7·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·11
7·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1·16

* These two values have been extrapolated at distances of ·25 and ·33 times the argument interval respectively, and if used must be accepted with caution.

TABLE (4).

Table of Deviates of Type I Curves Measured from the Mean in Terms of the Standard Deviation. $I_x(p, q) = 0·995$.

β_1.

β_2.	*0·00*	*0·01*	*0·03*	*0·05*	*0·10*	*0·15*	*0·20*	*0·30*	*0·40*	*0·50*	*0·60*	*0·70*	*0·80*	*0·90*	*1·00*	*1·20*	*1·40*	*1·60*	*1·80*	*2·00*	*2·20*	*2·40*	*2·60*	*2·80*	*3·00*
2·4	2·26	2·35^{+}	2·41	2·44*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·6	2·38	2·48	2·54	2·57	2·63	2·66	2·68*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2·8	2·49	2·58	2·64	2·68	2·73	2·77	2·80	2·83	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·0	2·58	2·66	2·72	2·76	2·82	2·86	2·89	2·93	2·95^{+}	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·2	—	—	—	—	2·89	2·93	2·96	3·01	3·04	3·06	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3·4	—	—	—	—	—	—	—	3·07	3·11	3·13	3·15^{-}	3·16	—	—	—	—	—	—	—	—	—	—	—	—	—
3·6	—	—	—	—	—	—	—	—	3·16	3·19	3·22	3·23	3·24	—	—	—	—	—	—	—	—	—	—	—	—
3·8	—	—	—	—	—	—	—	—	—	3·24	3·27	3·29	3·30	3·31	—	—	—	—	—	—	—	—	—	—	—
4·0	—	—	—	—	—	—	—	—	—	—	3·31	3·34	3·36	3·37	3·38	—	—	—	—	—	—	—	—	—	—
4·2	—	—	—	—	—	—	—	—	—	—	—	—	3·40	3·42	3·43	3·45^{+}	3·45^{-}	3·43	—	—	—	—	—	—	—
4·4	—	—	—	—	—	—	—	—	—	—	—	—	—	3·45^{+}	3·47	3·50^{+}	3·52	3·51	—	—	—	—	—	—	—
4·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·51	3·54	3·57	3·57	3·56	—	—	—	—	—	—
4·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·57	3·60	3·63	3·63	3·62	—	—	—	—	—
5·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·63	3·67	3·68	3·68	—	—	—	—	—
5·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·66	3·70	3·72	3·73	3·74	—	—	—	—
5·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·72	3·75^{+}	3·77	3·78	3·79	—	—	—
5·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·77	3·80	3·82	3·83	3·84	—	—
5·8	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·78	3·83	3·85^{+}	3·87	3·88	—	—
6·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·84	3·88	3·90	3·92	3·93	—
6·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·89	3·92	3·95^{-}	3·96	3·98
6·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·90	3·94	3·97	3·99	4·01
6·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·95^{-}	3·98	4·01	4·03
6·8	—	—	—	—	—		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·99	4·03	4·06
7·0	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4·00	4·04	4·07
7·2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4·05^{-}	4·08
7·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4·09
7·6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	4·10

* These two values have been extrapolated at distances of ·25 and ·33 times the argument interval respectively, and if used must be accepted with caution.

MUTUALLY CONSISTENT MULTIPLE REGRESSION SURFACES.

By Dr BURTON H. CAMP.

(1) *Introductory.* It is my purpose to study some of the properties of multiple correlation solids in which there are three variates only, x, y, and z, with special reference to the most probable forms of their regression surfaces and total regression curves. It will be desirable, incidentally, to develop certain properties of correlation surfaces in which the frequency is a function of two variates only.

As soon as one departs slightly from the normal case, in which all the regression surfaces are plane, and all the total regression curves are lines, it is quite customary to assume that the surfaces are polynomials of the second or higher degree; e.g., the regression of z on xy may be written:

$$z = g_0 + g_1 x + g_2 y + g_3 xy + g_4 x^2 + g_5 y^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1).$$

The values of the g's in a numerical case are then usually determined by least squares. It is frequently assumed that the total regression curves also are polynomials; e.g., the total regression of z on x may be assumed to be a polynomial of the second degree or higher in x. It takes but brief consideration, however, to assure one that if mathematical rigour is to be preserved such assumptions must not be made in random fashion. There are inter-relations among these regression surfaces and curves: if some are prescribed, others are determined. Further, even when the remaining regressions are not thus rigidly determined mathematically, they often are so determined save for a remote possibility which one is quite willing to assume non-existent. For instance, suppose the three-way frequency solid were required to become two-dimensional only, i.e., a mathematical surface. This would be a possibility that one would be unwilling to admit, for one would usually be sure that there was real variation with respect to all three of the variates, not merely with respect to two.

It becomes important, therefore, to separate out from among the various conceivable combinations that could be made a few simple, self-consistent groups. One such group might contain the above equation (1) as the regression of z on x, y. What simple assumptions could be made with regard to the other regression surfaces, and with regard to the total regression curves, which would not conflict with that equation or with each other? A few such cases will be described in Section (4). It is, however, rather the method used than the actual cases described which will be found useful. As a further example of the method, a detailed study is made of a frequency solid used by Dr L. Isserlis for which the data were supplied by Miss Elderton.

A study of the various cases leads to the following conclusion. If the regression surfaces are to be simple polynomials, but not planes, then a common form for the total regression to take will be:

$$y = \frac{\text{polynomial in } x}{\text{polynomial in } x} \qquad \text{.............................(2)},$$

and very often it will be much more complicated than this. Now it will be recognized that in almost every two-way correlation the frequencies are in reality the totals of an m-way correlation regression solid. The variation with regard to the other $(m-2)$ variables may not have been studied, and so the data may not be available to describe that solid, but it will be recognized that this variation and that solid do exist. So then, not to progress beyond the three-way solid for the moment, we may say that in general two-way regression is in reality a total regression for a three-way solid. Our study seems to imply, then, that the form (2) should occur frequently in the study of two-way correlation. This point is considered more fully in Section (6).

In Section (6) brief mention is made of the implications involved when we regard two-way correlation as one of the total correlations in a solid of more than three dimensions. Finally, further possible applications of the results are suggested, especially in the field of business forecasting.

This paper was begun while the author was at the Galton Laboratory. It is a pleasure to record his gratitude to Professor Karl Pearson for suggestions and kindly criticism.

(2) *Two Variates.* It is necessary first to formulate certain properties of correlation surfaces in which the frequency $n(x, y)$ is a function of two variates only, x and y. Most, if not all, of these have been observed before. They are easily derivable from the definitions. In making a correlation table, one divides the xy-plane into cells of area* $\Delta x \Delta y$. Then $n(x, y)$ is the frequency in the cell whose coordinates are (x, y). For our purposes it seems to be a trifle more convenient to use the language of integrals rather than of sums, and so, to be consistent, one may think of a correlation table as a plane slab of non-uniform density, and of $n(x, y)$ as the mass (frequency) "at the point" (x, y). Let n represent the total mass (total frequency). Then

$$\left.\begin{aligned} n(x) &= \int n(x, y)\,dy, \qquad n(y) = \int n(x, y)\,dx \\ n &= \int n(x)\,dx = \int n(y)\,dy = \iint n(x, y)\,dx\,dy \end{aligned}\right\} \qquad \text{.................(1)}.$$

If the origin be chosen at the "centre of gravity" of the slab, so that

$$\int n(x)\,x\,dx = \int n(y)\,y\,dy = 0 \qquad \text{.........................(2)},$$

then

$$\left.\begin{aligned} \sigma_x^2 &= \frac{1}{n}\int n(x)\,x^2\,dx = \frac{1}{n}\iint n(x, y)\,x^2\,dx\,dy \\ \sigma_y^2 &= \frac{1}{n}\int n(y)\,y^2\,dy = \frac{1}{n}\iint n(x, y)\,y^2\,dy\,dx \end{aligned}\right\} \qquad \text{.................(3)}.$$

---

* Some authors use the word cell to mean the area $y\Delta x$.

Let the units be so chosen that $\sigma_x = \sigma_y = 1$. As usual, the product moments and simple moments are

$$q_{x^a y^b} = \frac{1}{n}\iint n(x, y)\, x^a y^b\, dx\, dy,\quad \mu_{x^a} = q_{x^a y^0},\quad \mu_{y^b} = q_{x^0 y^b} \quad \ldots\ldots\ldots(4);$$

and in particular

$$r = q_{xy} = \frac{1}{n}\iint n(x, y)\, xy\, dx\, dy \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5).$$

$$\left.\begin{array}{ll} \text{Then} & \gamma(x) = \dfrac{1}{n(x)}\displaystyle\int n(x, y)\, y\, dy \text{ is the regression of } y \text{ on } x \\ \text{and} & \gamma'(y) = \dfrac{1}{n(y)}\displaystyle\int n(x, y)\, x\, dx \text{ is the regression of } x \text{ on } y \end{array}\right\} \ldots\ldots\ldots(6).$$

The partial ath moment about the mean of the column whose abscissa is x’ i.e., about the regression curve, is

$$\mu_{y^a . x} = \frac{1}{n(x)}\int n(x, y)(y - \gamma)^a\, dy \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(7),$$

and the corresponding partial ath moment about the x-axis is

$$\mu'_{y^a . x} = \frac{1}{n(x)}\int n(x, y)\, y^a\, dy \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8).$$

These partial moments are frequently spoken of as moments of arrays*; the word partial is used here because it is consistent with the language to be used with three-way solids. By equations (7) and (8)

$$\mu'_{y^a . x} = \mu_{y^a . x} + a\gamma\mu_{y^{a-1} . x} + \frac{a(a-1)}{2!}\gamma^2\mu_{y^{a-2} . x} + \ldots + \gamma^a, \text{ and } \mu_{y . x} = 0 \ldots(9).$$

It follows that, if $\gamma(x)$ be a polynomial of the pth degree in x, $\mu'_{y^a . x}$ is in general of the apth or higher degree. It will be of the apth degree even when the μ's, i.e., the moments about the regression curve, are constants, a condition which is sometimes described as "complete homoscedasticity in the x-direction." But it will also be of the apth degree often when these μ's are not constants, provided $\mu_{y^{a-k} . x}$ be of degree ap/k or less, for each $k < a$. Of course $\mu'_{y . x} = \gamma(x)$. We define $\mu_{x^b . y}$ and $\mu'_{x^b . y}$ in manner analogous to equations (7) and (8); $\mu'_{x . y} = \gamma'(y)$. As usual, let

$$\beta_1 = \mu_3^2/\sigma^6,\quad \beta_2 = \mu_4/\sigma^4 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10).$$

---

* Standard notation is followed very nearly. It is necessary in this paper to indicate the variable or variables of which each parameter is a function. This is done sometimes by the letters in the parenthesis following the parameter, sometimes by the letters following the dot in the subscript. It is also desirable to indicate the direction of the array to which the parameter pertains. This is done by the letter before the dot in the subscript, or by the whole subscript when it contains no dot. Thus $n(x, y)$ is a function of x and y; $n(x)$ is a function of x, and it might have been written $n_{y . x}$, since it is the total of a y-array and x is the coordinate of this array; σ_x is a constant and pertains to the marginal total which extends in the x direction. It will be observed that by "array" is meant either a column or a row.

The following statements, already well known, may now be easily established:

(*a*) If the regression, y on x, is linear, $\gamma(x)=rx$, $q_{xy}=r\mu_{x^2}=r$, $q_{x^2y}=r\mu_{x^3}$, $q_{x^3y^3}=r\mu_{x^4}$, etc.

(*b*) If also $\mu_{y^2.x}$ is constant, $\mu_{y^2.x}=1-r$; and if further $\mu_{y^3.x}$ is constant, $\mu_{y^3.x}=\mu_{y^3}-r^3\mu_{x^3}$; and if further $\mu_{y^4.x}$ is constant, $\mu_{y^4.x}=\mu_{y^4}-6r^2(1-r^2)-r^4\mu_{x^4}$.

(*c*) If both regressions are linear, and if $\mu_{y^2.x}$ and $\mu_{x^2.y}$ are constants,

$$\mu_{x^3}=\mu_{y^3}=0;$$

i.e., the β_1's of the marginal totals vanish (unless $r=0$, 1, or -1).

(*d*) If, in addition to (*c*), $\mu_{y^3.x}$ and $\mu_{x^3.y}$ are constants, they equal zero, and $\mu_{x^4}=\mu_{y^4}=3$, i.e., the marginal totals are of normal type ($\beta_1=\beta_2-3=0$), and the β_1 of each column and each row is zero (unless $r=0$, 1, or -1).

(*e*) If, in addition to (*d*), $\mu_{y^4.x}$ and $\mu_{x^4.y}$ are constants (homoscedasticity up to and including the β_2's), each column and each row is of normal type (unless $r=0$, 1, or -1).

It is desirable that the reader have these results in mind when forming a judgment as to the reasonableness of certain hypotheses to be made later. It will be noted that, if the marginal totals are not of symmetrical type ($\beta_1=0$), one must not assume linear regression and constant σ's for parallel arrays*. If the marginal totals are not of normal type ($\beta_1=\beta_2-3=0$), one must not assume linear regression and constant second and third partial moments for parallel arrays.

The reasoning which leads to these conclusions does not, however, compel similarly severe restrictions, so far as the first four moments are concerned, after one has departed from linear regression or from homoscedasticity. To show this I will now demonstrate (*e*), and then consider also the following case.

Case (*f*). Let both regressions be parabolic, and let the partial σ's of parallel arrays be constants.

Proof of (*e*). By hypothesis, $\mu_{y^4.x}=C$, $\mu_{x^4.y}=D$. By equations (8) and (9) the first of these statements may be written in the equivalent form:

$$\frac{1}{n(x)}\int n(x, y)\, y^4\, dy=C+6r^2x^2(1-r^2)+r^4x^4,$$

which is identically true for all values of x. We may therefore multiply both sides by $xn(x)\,dx/n$ and integrate, obtaining $q_{xy^4}=C+r^4\mu_{x^5}$. But by (*b*),

$$C=\mu_{y^4}-6r^2(1-r^2)-r^4\mu_{x^4}=3(1-r^2)^2,$$

by use of (*d*). Therefore

$$\frac{\mu_{y^4.x}}{\mu^2_{y^2.x}}=\frac{3(1-r^2)^2}{(1-r^2)^2}=3.\quad \text{Similarly } \frac{\mu_{x^4.y}}{\mu^2_{x^2.y}}=3.$$

Case (*f*). By hypothesis,

$$\gamma=A+Bx+Cx^2, \qquad \gamma'=A'+B'y+C'y^2,$$
$$\mu_{y^2.x}=D, \qquad \mu_{x^2.y}=D'.$$

---

* As stated before, by "array" is meant either row or column.

Applying equations (8) and (9) to each of the above, we have, from the two equations on the left:

$$\left.\begin{aligned}\frac{1}{n(x)}\int n(x,y)\,y\,dy &= A + Bx + Cx^2,\\ \frac{1}{n(x)}\int n(x,y)\,y^2\,dy &= D + (A + Bx + Cx^2)^2\end{aligned}\right\} \quad \ldots\ldots\ldots(11).$$

Multiply each of the identities in (11) by $x^i n(x)\,dx/n$ and integrate for $i = 0, 1, 2, \ldots$, but retain only those equations which involve moments of the fourth or lower degrees:

$$\left.\begin{aligned}0 &= A + C\\ r &= B + C\mu_{x^3}\\ q_{x^2y} &= A + B\mu_{x^3} + C\mu_{x^4}\\ 1 &= D + A^2 + B^2 + C^2\mu_{x^4} + 2AC + 2BC\mu_{x^3}\end{aligned}\right\} \quad .(12).$$

It is obvious that if the values of μ_{x^3} and μ_{x^4}, in fact even of r and q_{x^2y} also, were preassigned arbitrarily, it would still be possible usually to find values for A, B, C, and D. Similar reasoning applies to the letters A', B', C', and D'. As far as our previous reasoning has led us, therefore, there is nothing to prevent an assumption of parabolic regression and constant partial σ's, no matter to what types of frequency distribution the marginal totals may belong.

(3) *Three Variates.* In constructing a multiple correlation solid of three dimensions, one divides ordinary space into cells* of volume $\Delta x \Delta y \Delta z$. Let $n(x, y, z)$ be the frequency in the cell whose coordinates are (x, y, z); or, if one wishes to use the language of integrals rather than of sums, one may think of a non-homogeneous solid and of $n(x, y, z)$ as the density at the point (x, y, z). The total frequency of the z column whose coordinates are (x, y) may be written, in conformity with the notation of the previous section, $n(x, y)$, or $n_{z.xy}$, and

$$n(x,y) = \int n(x,y,z)\,dz.$$

The total frequency of the slab parallel to the yz plane whose coordinate is x is

$$n(x) = n_{zy.x} = \int n(x,y)\,dy = \iint n(x,y,z)\,dy\,dz \quad \ldots\ldots\ldots\ldots\ldots(1).$$

The total frequency of the solid is

$$n = \int n(x)\,dx = \iint n(x,y)\,dx\,dy = \iiint n(x,y,z)\,dx\,dy\,dz \quad \ldots\ldots\ldots\ldots(2).$$

Let the origin be chosen at the "centre of gravity" of this solid so that

$$\iiint n(x,y,z)\,x\,dx\,dy\,dz = \int n(x)\,x\,dx = \int n(y)\,y\,dy = \int n(z)\,z\,dz = 0 \quad \ldots(3).$$

Then

$$\sigma_x^2 = \frac{1}{n}\iiint x^2\,n(x,y,z)\,dx\,dy\,dz = \frac{1}{n}\iint x^2\,n(x,y)\,dx\,dy = \frac{1}{n}\int x^2\,n(x)\,dx \quad \ldots(4).$$

---

* Some authors define cell so that it has the volume $z\,\Delta x\,\Delta y$.

Let the units be so chosen that

$$\sigma_x = \sigma_y = \sigma_z = 1 \quad \dots\dots\dots\dots\dots\dots\dots\dots(5).$$

As usual, the product moments and simple moments are denoted by

$$q_{x^a y^b z^c} = \frac{1}{n} \iiint n\,(x, y, z)\, x^a y^b z^c\, dx\, dy\, dz,$$

and in particular

$$\mu_{x^a} = q_{x^a y^0 z^0}$$

is the ath moment of the marginal total parallel to the x axis; $r_{xy} = q_{xyz^0}$ is the total correlation between x and y, i.e., the simple two-way correlation between the totals of the z columns. Let $g\,(x, y)$ be the mean of the z column whose coordinates are (x, y), then:

$$g\,(x, y) = \frac{1}{n\,(x, y)} \int n\,(x, y, z)\, z\, dz \quad \dots\dots\dots\dots\dots\dots(6).$$

Regarded as an equation in x and y this defines the locus of the mean points of all z columns and is therefore the equation of the regression surface, z on xy. Similarly, the regression y on xz is

$$h\,(x, z) = \frac{1}{n\,(x, z)} \int n\,(x, y, z)\, y\, dy \quad \dots\dots\dots\dots\dots\dots(7),$$

and of x on yz is

$$k\,(y, z) = \frac{1}{n\,(y, z)} \int n\,(x, y, z)\, x\, dx \quad \dots\dots\dots\dots\dots\dots(8).$$

The section of the $g\,(x, y)$ surface obtained by fixing x is the partial regression of z on y for the fixed x. It is the two-way regression curve, z on y, in the slab whose coordinate is x. No special notation is required; it is only necessary to regard x as fixed in (6). Equation (6) thus may be thought of as the equation either of a regression surface or of a partial regression curve. The mean of that marginal total of this slab which extends in the z direction is to be designated by $\alpha\,(x)$, and the equation $z = \alpha\,(x)$ is the equation of the total regression, z on x. There are six total regressions:

$$\left.\begin{array}{ll}
z \text{ on } x: & \alpha\,(x) = \dfrac{1}{n\,(x)} \displaystyle\int n\,(x, y)\, g\,(x, y)\, dy = \dfrac{1}{n\,(x)} \iint n\,(x, y, z)\, z\, dy\, dz \\[2ex]
x \text{ on } z: & \alpha'\,(z) = \dfrac{1}{n\,(z)} \displaystyle\int n\,(z, y)\, k\,(y, z)\, dy = \dfrac{1}{n\,(z)} \iint n\,(x, y, z)\, x\, dy\, dx \\[2ex]
z \text{ on } y: & \beta\,(y) = \dfrac{1}{n\,(y)} \displaystyle\int n\,(y, x)\, g\,(x, y)\, dx = \dfrac{1}{n\,(y)} \iint n\,(x, y, z)\, z\, dx\, dz \\[2ex]
y \text{ on } z: & \beta'\,(z) = \dfrac{1}{n\,(z)} \displaystyle\int n\,(x, z)\, h\,(x, z)\, dx = \dfrac{1}{n\,(z)} \iint n\,(x, y, z)\, y\, dx\, dy \\[2ex]
y \text{ on } x: & \gamma\,(x) = \dfrac{1}{n\,(x)} \displaystyle\int n\,(x, z)\, h\,(x, z)\, dz = \dfrac{1}{n\,(x)} \iint n\,(x, y, z)\, y\, dy\, dz \\[2ex]
x \text{ on } y: & \gamma'\,(y) = \dfrac{1}{n\,(y)} \displaystyle\int n\,(y, z)\, k\,(y, z)\, dz = \dfrac{1}{n\,(y)} \iint n\,(x, y, z)\, x\, dx\, dz
\end{array}\right\} \dots(9).$$

The notation is in agreement with that of Section (2). For example, if we integrate $n\,(x, y, z)$ with respect to z, we get $n\,(x, y)$, a set of totals which define a two-way correlation table like that of Section (2). By equation (6) of Section (2) we have also

$$\gamma'(y) = \frac{1}{n(y)}\int n(x, y)\, x\, dx = \frac{1}{n(y)}\iint n(x, y, z)\, x\, dx\, dz,$$

which is in agreement with the last equation of (9) above. Finally it is necessary to write equations for the "partial product and simple moments about the coordinate axes." These are the ordinary two-way moments about the coordinate axes in the correlation tables defined by the slabs of our solid. Thus, in the slab whose coordinate is x, the partial product moment of order ab is

$$q'_{y^a z^b . x} = \frac{1}{n(x)}\iint n(x, y, z)\, y^a z^b\, dy\, dz \text{ .... (10).}$$

When $b = 0$ in this, it becomes a partial simple moment:

$$q'_{y^a z^0 . x} = \mu'_{y^a . x} \text{ .... (11),}$$

and when also $a = 1$, it becomes the first partial simple moment, which, regarded as a function of x, is the total regression, y on x:

$$\mu'_{y . x} = \gamma(x) \text{ .... (12).}$$

(4) *Relations between Regression Surfaces and Curves.* It is clear from the formulae of Section (3) that there exist relations between the regression surfaces and the total regression curves of a solid of frequency. In developing these relations we have the option of two alternative courses. Either we may assume simple expressions for our total regression curves, and find the resultant (usually more complicated) forms for our regression surfaces; or we may assume simple expressions for our regression surfaces, and find the resultant (likewise usually more complicated) forms for our total regression curves. Which assumption is likely to prove the more useful? We may put the question in another form which will suggest the answer. As pointed out before, the regression surface, say $g(x, y)$, is, when one of the variables, say y, is fixed, a partial regression curve, z on x. Now, which is simpler, this partial regression, or the total regression of z on x? Usually the former, the partial regression, is simpler, because, when y is fixed, one is dealing with a more homogeneous population than when y is allowed to vary. An example, to be employed numerically later, will illustrate this remark. Let x measure the height, y the weight, and z the age of an individual. The relation between height and weight is expected to be simpler for that portion of the population all members of which have the same age than for the whole group. The reader should not confuse this question with the question of sampling. The fact that in any finite group the number of individuals whose age was the same would be small, and therefore large discrepancies from the ideal relation apt to occur, has nothing to do with the question. We are concerned here with the ideal relation itself, not with the sampling variation from it. We may think of the population as infinite, or we may state the example in another way: Select at random a million Englishmen; also a million Englishmen whose age is twenty-one. The relation between height and weight should be simpler in the second case than in

the first—unless it be linear in both. We choose, then, the second of the alternatives, and state, at first without demonstration, the following theorems and corollaries.

Theorem I. Let all the partial regressions be linear functions, such as:

$$g(x, y) = g_{00} + g_{01}y + g_{10}x + g_{11}xy.$$

Then all the total regressions are of the form:

$$\alpha(x) = \frac{\text{parabola}^{*} \text{ in } x}{\text{parabola in } x}.$$

Corollary. If also the regression surfaces be linear—so that in particular $g_{11} = 0$, then all the total regressions are linear.

This is the well-known all linear case. To obtain this sort of total regression, it is not necessary to assume a normal solid of frequency or homoscedasticity, provided one assumes linear regression surfaces.

Theorem II. Let all the partial regressions be linear, except that $g(x, y)$ shall be parabolic in x, i.e.,

$$g(x, y) = g_{00} + g_{01}y + g_{10}x + g_{11}xy + g_{20}x^2 + g_{21}x^2y.$$

Then in general† the total regressions will have the forms:

$$\begin{aligned}
&(\text{Cubic in } x)\,\alpha + (\text{cubic in } x) = 0,\\
&(\text{Parabola in } z)\,\alpha' + (\text{parabola in } z) = 0,\\
&(\text{Cubic in } y)\,\beta^2 + (\text{cubic in } y)\,\beta + (\text{cubic or higher in } y) = 0,\\
&(\text{Parabola in } z)\,\beta' + (\text{parabola in } z) = 0,\\
&(\text{Cubic in } x)\,\gamma + (\text{cubic in } x) = 0,\\
&(\text{Parabola in } y)\,\gamma'^2 + (\text{parabola in } y)\,\gamma' + (\text{parabola or higher in } y) = 0.
\end{aligned}$$

Corollary. If the product terms in the regression surface formulae be omitted, and if the partial σ_x of the slab perpendicular to the y axis be constant ($\mu_{x^2.y}$ = constant in y); then in general the total regressions have the forms:

$$\begin{aligned}
&\alpha = \text{parabola}^{\ddagger},\\
&\alpha' = \text{line},\\
&\beta^2 + (\text{line})\,\beta + (\text{parabola}) = 0,\\
&\beta' = \text{line},\\
&\gamma = \text{parabola},\\
&\gamma'^2 + (\text{constant})\,\gamma' + (\text{line}) = 0.
\end{aligned}$$

---

[*] I shall use the words, parabola, cubic, etc., in x, to denote a polynomial of the 2nd, 3rd, etc., degrees in x; and the words, parabola or higher in x, to denote a polynomial in x of at least the second

[†] I.e., except for specialized cases, one of which is given in the corollary following, where these forms are simplified.

[‡] Parabola in x. Since α is a function of x only, this is obvious, and so the words "in x" may be omitted. Hereafter, in like cases, the independent variable will not be written.

Theorem III. Let all the partials be linear except that $g(x, y)$ and $k(z, y)$ shall be parabolic in x and z respectively. Then in general the forms for the total regressions are:

$$\begin{aligned}
&(\text{Cubic})\,\alpha + (\text{cubic}) = 0,\\
&(\text{Cubic})\,\alpha' + (\text{cubic}) = 0,\\
&(\text{Cubic})\,\beta^4 + (\text{cubic})\,\beta^3 + (\text{cubic or higher})\,\beta^2 + (\text{cubic or higher})\,\beta\\
&\qquad\qquad + (\text{cubic or higher}) = 0,\\
&(\text{Cubic})\,\beta' + (\text{cubic}) = 0,\\
&(\text{Cubic})\,\gamma + (\text{cubic}) = 0,
\end{aligned}$$

and γ' has the same form as β.

Corollary. If the product terms in the regression surfaces be dropped, and if the partial σ's of the slabs perpendicular to the y axis be constants ($\mu_{z^2 . y}$, $\mu_{x^2 . y}$ equal constants in y); then in general the total regressions have the forms:

$$\alpha,\ \alpha',\ \beta',\ \gamma = \text{parabolae};\ \beta \text{ and } \gamma' \text{ have the same form namely:}$$

$$\beta^4 + (\text{constant})\,\beta^3 + (\text{line})\,\beta^2 + (\text{line})\,\beta + (\text{parabola}) = 0.$$

Theorem IV. Let all the partial regressions be parabolic, except that $h(x, z)$ and $k(y, z)$ shall be linear in x and y respectively. Then in general the forms for the total regressions are:

$$\begin{aligned}
&(\text{Quartic})\,\alpha^4 + (\text{quartic})\,\alpha^3 + (\text{quartic or higher})\,\alpha^2 + (\text{quartic or higher})\,\alpha\\
&\qquad\qquad + (\text{quartic or higher}) = 0,
\end{aligned}$$

β has the same form as α,

$$(\text{Quartic})\,\alpha' + (\text{quartic}) = 0,$$

β' has the same form as α',

$$\begin{aligned}
&(\text{Quintic})\,\gamma^4 + (\text{quintic})\,\gamma^3 + (\text{quintic or higher})\,\gamma^2 + (\text{quintic or higher})\,\gamma\\
&\qquad\qquad + (\text{quintic or higher}) = 0,
\end{aligned}$$

and γ' has the same form as γ.

Theorem V. Let all the partial regressions be parabolic. Then in general the forms for the total regressions are:

$$\begin{aligned}
&(\text{Sextic})\,\alpha^4 + (\text{sextic})\,\alpha^3 + (\text{sextic or higher})\,\alpha^2 + (\text{sextic or higher})\,\alpha\\
&\qquad\qquad + (\text{sextic or higher}) = 0,
\end{aligned}$$

and similar equations for the other variables.

The number of such theorems could be increased at pleasure, but these are sufficient to give the reader a general idea of what to expect. They are not intended to be sufficient to include many of the cases which will occur in practice. The method of derivation will now be illustrated in the proof of one of these theorems, and by application to a practical case. Then some comment will be made on the general inferences to be drawn.

Proof of Theorem III. By hypothesis the regression surfaces have the forms:

$$\left.\begin{aligned}
g(x, y) &= g_{00} + g_{01}y + g_{10}x + g_{11}xy + g_{20}x^2 + g_{21}x^2y\\
h(x, z) &= h_{00} + h_{01}z + h_{10}x + h_{11}xz\\
k(y, z) &= k_{00} + k_{01}z + k_{10}y + k_{11}yz + k_{02}z^2 + k_{12}yz^2
\end{aligned}\right\} \quad .(1).$$

We first find the totals by the use of equations (9) of Section (3). The first of the above equations holds identically for all values of x and y, and we may multiply through by $n(x,y)\,dy/n(x)$ and integrate with respect to y, obtaining the following identity in x:

$$\alpha(x) = g_{00} + g_{10}x + g_{20}x^2 + (g_{01} + g_{21}x^2)\frac{1}{n(x)}\int n(x,y)\,y\,dy.$$

But $$\frac{1}{n(x)}\int n(x,y)\,y\,dy = \frac{1}{n(x)}\iint n(x,y,z)\,y\,dy\,dz = \gamma(x).$$

So, for all values of x identically:

$$\left.\begin{aligned}\alpha &= g_{00} + g_{10}x + g_{20}x^2 + (g_{01} + g_{11}x + g_{21}x^2)\gamma\\ \gamma &= h_{00} + h_{10}x + (h_{01} + h_{11}x)\,\alpha\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots(2),$$

and, similarly,

$$\left.\begin{aligned}\alpha' &= k_{00} + k_{01}z + k_{02}z^2 + (k_{01} + k_{11}z + k_{12}z^2)\,\beta'\\ \beta' &= h_{00} + h_{01}z + (h_{10} + h_{11}z)\,\alpha'\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots(3),$$

and, using equation (11) of Section (3),

$$\left.\begin{aligned}\beta &= g_{00} + g_{01}y + (g_{10} + g_{11}y)\,\gamma' + (g_{20} + g_{21}y)\,\mu'_{x^2.y}\\ \gamma' &= k_{00} + k_{10}y + (k_{01} + k_{11}y)\,\beta + (k_{02} + k_{12}y)\,\mu'_{z^2.y}\end{aligned}\right\} \quad \ldots\ldots(4).$$

We may now solve each pair of equations for the values of the Greek letters. By (2)

$$\alpha = g_{00} + g_{10}x + g_{20}x^2 + (g_{01} + g_{11}x + g_{21}x^2)\,\{h_{00} + h_{10}x + (h_{01} + h_{11}x)\,\alpha\},$$

$$\begin{aligned}&\alpha\,\{1 - (g_{01} + g_{11}x + g_{21}x^2)(h_{01} + h_{11}x)\}\\ &\qquad = g_{00} + g_{10}x + g_{20}x^2 + (g_{01} + g_{11}x + g_{21}x^2)(h_{00} + h_{10}x):\end{aligned}$$

i.e., (cubic in x) α + (cubic in x) $= 0$, in general, save for the vanishing of certain coefficients, as for example as stated in the corollary. From (2) we may also derive:

$$\gamma = h_{00} + h_{10}x + (h_{01} + h_{11}x)\,\{g_{00} + g_{10}x + g_{20}x^2 + (g_{01} + g_{11}x + g_{21}x^2)\,\gamma\},$$

$$\begin{aligned}&\gamma\,\{1 - (h_{01} + h_{11}x)(g_{01} + g_{11}x + g_{21}x^2)\}\\ &\qquad = h_{00} + h_{10}x + (h_{01} + h_{11}x)(g_{00} + g_{10}x + g_{20}x^2);\end{aligned}$$

i.e., (cubic in x) γ + (cubic in x) $= 0$. Equations (3) are exactly similar to equations (2) and so the expressions for α' and β' will be similar to those for α and γ. Before solving equations (4), we write, by equations (9) of Section (2),

$$\mu'_{x^2.y} = \mu_{x^2.y} + (\gamma')^2, \qquad \mu'_{z^2.y} = \mu_{z^2.y} + \beta^2.$$

Then (4) give:

$$\begin{aligned}\beta &= g_{00} + g_{01}y + (g_{10} + g_{11}y)\,\{k_{00} + k_{10}y + (k_{01} + k_{11}y)\,\beta + (k_{02} + k_{12}y)(\mu_{z^2.y} + \beta^2)\}\\ &\quad + (g_{20} + g_{21}y)\,[\mu_{x^2.y} + \{k_{00} + k_{10}y + (k_{01} + k_{11}y)\,\beta + (k_{02} + k_{12}y)(\mu_{z^2.y} + \beta^2)\}^2];\end{aligned}$$

i.e.,

$$\begin{aligned}\beta &= \text{line} + \text{line}\,\{\text{line} + (\text{line})\,\beta + (\text{line})\,\mu_{z^2.y} + (\text{line})\,\beta^2\}\\ &\quad + (\text{line})\,\{\mu_{x^2.y} + (\text{line})^2 + (\text{line})^2\beta^2 + (\text{line})^2\mu_{z^2.y} + (\text{line})^2\beta^4\\ &\quad + 2\,[(\text{line})(\text{line})\,\beta + (\text{line})(\text{line})\,\mu_{z^2.y} + (\text{line})(\text{line})\,\beta^2\\ &\quad + (\text{line})(\text{line})\,\beta\mu_{z^2.y} + (\text{line})(\text{line})\,\beta^3 + (\text{line})(\text{line})\,\mu_{z^2.y}\beta^2]\}.\end{aligned}$$

Hence:

$$\begin{aligned}&\beta^4(\text{cubic}) + \beta^3(\text{cubic}) + \beta^2(\text{cubic} + \text{cubic}\,\mu_{z^2.y})\\ &\qquad + \beta\,(\text{cubic} + \text{cubic}\,\mu_{z^2.y}) + (\text{line}\,\mu_{x^2.y} + \text{cubic}\,\mu_{z^2.y} + \text{cubic}) = 0.\end{aligned}$$

This equation verifies the statement given in the theorem regarding β. If the partial moments $\mu_{z^2.y}$ and $\mu_{x^2.y}$ are constants in y, the words "or higher" in that statement may be deleted; otherwise they remain. It is obvious that the forms of β and γ' must be similar.

(5) *Elderton-Isserlis Example.* In the second[*] of two papers entitled, "On the Partial Correlation Ratio," Dr L. Isserlis has discussed some data prepared by Miss E. M. Elderton, in which the height (x), weight (y), and age (z) of a group of individuals are correlated. The units and origin chosen for these variables in that paper are the same as those chosen by me in this one, and so it is easy to write in our notation the conditions of Isserlis. He finds that

$$g(x, y) = g_{00} + g_{01}y + x(g_{10} + g_{11}y) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (A)$$

is a good fit to the regression of z on xy, and he states that Soper has found a good fit also to the regression of y on xz by using

$$h(x, z) = h_{00} + h_{01}z + h_{02}z^2 + x(h_{10} + h_{11}z + h_{12}z^2) \quad \ldots\ldots\ldots\ldots\ldots (B).$$

No attempt is made to find the form of $k(y, z)$. It will be our object to discover whether, consistently with (A) and (B), and with certain obvious restrictions on the total regressions to be mentioned directly, it is possible that $k(y, z)$ should be a polynomial of as low a degree as the second in y and z separately. So let

$$k(y, z) = k_{00} + k_{01}z + k_{02}z^2 + y(k_{10} + k_{11}z + k_{12}z^2) + y^2(k_{20} + k_{21}z + k_{22}z^2) \quad \ldots (C).$$

A glance at the regression curves pictured on pages 53—56 of Isserlis's paper is sufficient to convince one that the following may also be safely assumed: (a) the total regression of z on x, $\alpha(x)$, is not a conic section or a straight line; (b) the same is true of the total regression of z on y, $\beta(y)$; (c) the total regression of x on y, $\gamma'(y)$, is not a constant; and (d) the same is true of the total regression of y on x, $\gamma(x)$. With regard to the remaining two regressions, they appear to be nearly linear, and although this may be an illusion, we shall so assume: (e) the total regression of y on z, $\beta'(z)$, is a line; and (f) the total regression of x on z, $\alpha'(z)$, is a line. Our proposed problem might, therefore, be stated as follows. Is it possible that $k(y, z)$ shall be a polynomial as simple as (C) and at the same time that α' and β' shall be linear, without contradicting at least one of the necessary hypotheses, (A), (B), (a), (b), (c), and (d)? It would seem to be worth while undertaking such a discussion before proceeding further with the numerical problem of fitting the regression surfaces of this solid.

Starting from (A), (B), and (C), we may write our total regressions:

$$\left.\begin{aligned} \alpha\ (x) &= g_{00} + g_{10}x + (g_{01} + g_{11}x)\gamma \\ \gamma\ (x) &= h_{00} + h_{10}x + (h_{01} + h_{11}x)\alpha + (h_{02} + h_{12}x)\mu'_{z^2.x} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1),$$

$$\left.\begin{aligned} \alpha'\ (z) &= k_{00} + k_{01}z + k_{02}z^2 + (k_{10} + k_{11}z + k_{12}z^2)\beta' + (k_{20} + k_{21}z + k_{22}z^2)\mu'_{y^2.z} \\ \beta'\ (z) &= h_{00} + h_{01}z + h_{02}z^2 + (h_{10} + h_{11}z + h_{12}z^2)\alpha' \end{aligned}\right\} \quad (2),$$

$$\left.\begin{aligned} \beta\ (y) &= g_{00} + g_{01}y + (g_{10} + g_{11}y)\gamma' \\ \gamma'\ (y) &= (k_{00} + k_{10}y + k_{20}y^2) + (k_{01} + k_{11}y + k_{21}y^2)\beta + (k_{02} + k_{12}y + k_{22}y^2)\mu'_{z^2.y} \end{aligned}\right\} \quad (3),$$

where $\mu'_{z^2.x} = \mu_{z^2.x} + \alpha^2$, $\mu'_{y^2.z} = \mu_{y^2.z} + \beta'^2$, $\mu'_{z^2.y} = \mu_{z^2.y} + \beta^2$ ………(4).

---

[*] *Biometrika*, Vol. XI. pp. 50—66.

First solve (2):

$$\begin{aligned}&\alpha'^2(k_{20}+k_{21}z+k_{22}z^2)(h_{10}+h_{11}z+h_{12}z^2)^2\\ &\quad+\alpha'\{2(k_{20}+k_{21}z+k_{22}z^2)(h_{00}+h_{01}z+h_{02}z^2)(h_{10}+h_{11}z+h_{12}z^2)\\ &\quad+(k_{10}+k_{11}z+k_{12}z^2)(h_{10}+h_{11}z+h_{12}z^2)-1\}+(k_{20}+k_{21}z+k_{22}z^2)\{\mu_{y^2.z}\\ &\quad+(h_{00}+h_{01}z+h_{02}z^2)^2\}+(k_{10}+k_{11}z+k_{12}z^2)(h_{00}+h_{01}z+h_{02}z^2)+k_{00}+k_{01}z+k_{02}z^2=0\\ &\hspace{30em}\text{......(5)},\end{aligned}$$

$$\begin{aligned}&\beta'^2(h_{10}+h_{11}z+h_{12}z^2)(k_{20}+k_{21}z+k_{22}z^2)+\beta'\{(h_{10}+h_{11}z+h_{12}z^2)(k_{10}+k_{11}z+k_{12}z^2)-1\}\\ &\quad+h_{00}+h_{01}z+h_{02}z^2+(h_{10}+h_{11}z+h_{12}z^2)(k_{00}+k_{01}z+k_{02}z^2)=0\ \text{..................(6)}.\end{aligned}$$

By hypothesis, (5) and (6) must be straight lines. It is therefore necessary that either:

(i) $k_{20}+k_{21}z+k_{22}z^2=0$ for all values of z, and not (ii); or

(ii) $h_{10}+h_{11}z+h_{12}z^2=0$ for all values of z, and not (i); or

(iii) both (i) and (ii).

Before solving the other equations (1) and (3), it is convenient to pursue each of the above three possibilities.

Suppose (i). Then $k_{20}=k_{21}=k_{22}=0$, and equation (5) becomes:

$$\begin{aligned}&\alpha'\{(k_{10}+k_{11}z+k_{12}z^2)(h_{10}+h_{11}z+h_{12}z^2)-1\}\\ &\qquad\qquad+(k_{10}+k_{11}z+k_{12}z^2)(h_{00}+h_{01}z+h_{02}z^2)+k_{00}+k_{01}z+k_{02}z^2=0.\end{aligned}$$

For this to be a line it is necessary that:

$$k_{12}h_{12}=0,\ \ k_{11}h_{12}+k_{12}h_{11}=0,\ \ k_{10}h_{12}+k_{11}h_{11}+k_{12}h_{10}=0,\ \ k_{10}h_{11}+k_{11}h_{10}=0\ \text{...(7)},$$

and that

$$k_{12}h_{02}=0,\ \ k_{11}h_{02}+k_{12}h_{01}=0,\ \ k_{02}+k_{10}h_{02}+k_{11}h_{01}+k_{12}h_{00}=0\ \text{.........(8)}.$$

Equation (6) now becomes:

$$\begin{aligned}&\beta'\{(h_{10}+h_{11}z+h_{12}z^2)(k_{20}+k_{21}z+k_{22}z^2)-1\}\\ &\qquad\qquad+h_{00}+h_{01}z+h_{02}z^2+(h_{10}+h_{11}z+h_{12}z^2)(k_{00}+k_{01}z+k_{02}z^2)=0,\end{aligned}$$

and for this to be a straight line it is necessary that, in addition to (7) and (8),

$$k_{02}h_{12}=0,\ \ k_{02}h_{11}+k_{01}h_{12}=0,\ \ h_{02}+k_{00}h_{12}+k_{01}h_{11}+k_{02}h_{10}=0\ \text{.........(9)}.$$

Consider the inferences to be drawn from (7), (8), and (9). It is necessary by (7) that, either

(i*a*) $k_{12}=0$, and $h_{12}\neq 0$; or

(i*b*) $k_{12}=0$, and $h_{12}=0$; or

(i*c*) $k_{12}\neq 0$, and $h_{12}=0$.

First suppose (i*a*). Then (7), (8), and (9) become:

$$k_{11}=k_{10}=k_{02}=k_{01}=h_{02}+k_{00}h_{12}=0\ \text{.....................(10)}.$$

Condition (i) and equation (10) placed in equation (3) lead to the conclusion that $\gamma'(y)=k_{00}$, contradicting hypothesis (*c*).

Next suppose (ib). We have $k_{12}=k_{20}=k_{21}=k_{22}=h_{12}=0$, and equations (7), (8), and (9) become:

$$k_{11}h_{11}=k_{10}h_{11}+k_{11}h_{10}=k_{11}h_{02}=k_{02}+k_{10}h_{02}+k_{11}h_{01}=k_{02}h_{11}=h_{02}+k_{01}h_{11}+k_{02}h_{10}=0\ldots(11),$$

and it becomes necessary to consider three further cases:

Case (ib_1): $k_{11}=0,\ h_{11}\neq 0$; *Case* (ib_2): $k_{11}=0,\ h_{11}=0$; *Case* (ib_3): $k_{11}\neq 0,\ h_{11}=0$.

Suppose (ib_1). Then $k_{10}=0,\ \ k_{02}=0,\ \ h_{02}+k_{01}h_{11}=0$,

and, by (3),
$$\beta=g_{00}+g_{01}y+(g_{10}+g_{11}y)(k_{00}+k_{01}\beta),$$
which contradicts hypothesis (b).

Suppose (ib_2). Then $k_{02}+k_{10}h_{02}=0,\ \ h_{02}+k_{02}h_{10}=0$.

This supposition does not lead to a direct contradiction of the hypothesis, and will be considered further, after disposing of the other possibilities, all of which we shall find do lead to contradictions.

Suppose (ib_3). Then $h_{10}=h_{02}=k_{02}+k_{11}h_{01}=0$,

and, by (1),
$$\alpha=g_{00}+g_{10}x+(g_{01}+g_{11}x)(h_{00}+h_{01}\alpha),$$
contradicting hypothesis (a).

Now suppose (ic). We have

$$k_{12}\neq 0,\ \ h_{12}=k_{20}=k_{21}=k_{22}=0,$$

and so, by (7), (8), and (9), the following:

$$h_{11}=h_{10}=h_{02}=h_{01}=k_{02}+k_{12}h_{00}=0 \ \ldots\ldots\ldots\ldots\ldots\ldots \ (12).$$

By (1), then, $\gamma(x)=h_{00}$, contradicting hypothesis (d).

Suppose (ii). Since $h_{10}=h_{11}=h_{12}=0$,

equations (5) and (6) become:

$$-\alpha'+(k_{20}+k_{21}z+k_{22}z^2)\{\mu_{y^2.z}+(h_{00}+h_{01}z+h_{02}z^2)^2\}$$
$$+(k_{10}+k_{11}z+k_{12}z^2)(h_{00}+h_{01}z+h_{02}z^2)+k_{00}+k_{01}z+k_{02}z^2=0,$$

and
$$\beta'-h_{00}-h_{01}z-h_{02}z^2=0.$$

For these to be lines it is necessary that $h_{02}=0$. Then, by (1),

$$\alpha=g_{00}+g_{10}x+(g_{01}+g_{11}x)(h_{00}+h_{01}\alpha),$$

contradicting hypothesis (a).

Suppose (iii). The argument against (ii) holds here, and eliminates this case also. Let us return then to the only case which has not been proved impossible:

Case (ib_2). We had

$$k_{11}=k_{12}=k_{20}=k_{21}=k_{22}=h_{11}=h_{12}=0;$$

and
$$k_{02}+k_{10}h_{02}=0,\ \ h_{02}+k_{02}h_{10}=0.$$

From the last two of these equations we may derive:

$$h_{02}(1-k_{10}h_{10})=0, \text{ and } k_{02}(1-k_{10}h_{10})=0;$$

whence, either $k_{10}h_{10}=1$, or else $h_{02}=k_{02}=0$, but in the latter case, by (3),

$$\beta = g_{00}+g_{01}y+(g_{10}+g_{11}y)\,\{(k_{00}+k_{10}y)+k_{01}\beta\},$$

contradicting hypothesis (*b*). So we should add to the above

$$k_{10}h_{10}=1 \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(13).$$

Our conditions involve the regression surfaces:

$$\left.\begin{aligned} g(x,y) &= g_{00}+g_{01}y+x(g_{10}+g_{11}y)\\ h(x,z) &= h_{00}+h_{01}z+h_{02}z^2\ +xh_{10}\\ k(y,z) &= k_{00}+k_{01}z+k_{02}z^2\ +yk_{10}\end{aligned}\right\} \quad \dots\dots\dots\dots\dots\dots(14).$$

Now it is probable that the regression $h(x, z)$ cannot be fitted well by the expression in (14), for if it could be, Soper would probably not have used so many terms; but, in addition to this practical difficulty, the relation (13) imposes a restriction so severe as to make the form of $k(y, z)$ given above highly improbable. That is, if h_{10} has been determined before the k's, then there are left but three arbitrary constants in $k(y, z)$. So it would probably not pay to pursue this case (i b_2) further by numerical work. It would seem necessary, therefore, to generalize the conditions laid down at the outset of this section if we are to fit regression surfaces to this solid.

(6) The following conclusions may be drawn from the developments of the preceding pages.

(*a*) It is known, as stated in Section (2), that, in a two-way correlation table, severe restrictions cannot be placed on the regression curves without restricting the type of frequency distribution to which the marginal totals may belong, and that, if one knew in advance, as often happens, that the frequency distributions of the marginal totals were not of such restricted type, it would be inconsistent to restrict too severely the regression curves. It has now been shown that there are also, not exactly similar but similarly important, inter-relations in a three-way table. If the total regression curves are determined first, as they usually would be, and if they are not of the types described in Section (4), it would be inconsistent to suppose the regression surfaces such simple polynomials as are there described*.

(*b*) It has been found that the simplest assumptions frequently lead to total regressions at least as complicated in form as

$$y=\frac{\text{polynomial in } x}{\text{polynomial in } x}.$$

---

* This does not mean, of course, that it would not then be permissible in practice to try to fit the data with those simple polynomials. In so doing, however, one should recognize that an approximation is being made: the simple polynomial forms used should be thought of as approximations to the more complex forms strictly required. It is useful to know what are the mathematically self consistent types rigorously demanded by a given problem, but this knowledge does not prejudice the conscious use of any approximate substitutes for them which may be practically convenient. Likewise, the remark in Section (1), that hyperbolic forms (equation 2) should be thought of as occurring frequently in two-way tables, does not exclude the use of polynomials in two-way tables, at least over a moderate range of the independent variate; but the hyperbolic forms have a decided theoretical advantage, and so, except that they may increase the burden of computation, they should be regarded as preferable.

It seems probable that regressions of this type should be regarded as of common occurrence, and that when in a two-way table regression is not linear a regression of this type should be one of the first tried. For every two-way correlation table may be thought of as really a table of totals from a three-way table, as pointed out in Section (1). Another argument for two-way regressions of this form is that one of Narumi's simplest surfaces[*] gives regressions of the hyperbolic form:

$$y = \frac{a}{cx + d}.$$

(*c*) The reader will have been aware that the argument in (*b*) leads further than was indicated. If we are to infer the common forms of two-way regression curves by regarding them as the total regression curves of a three-way frequency solid, we may also regard them as the total regression curves of a four-way frequency solid, and proceed by making simple hypotheses as to its three-way regression solids; and so on, to any number of dimensions. It is undesirable to lengthen this paper by a detailed study of four-way solids, but the author has in fact carried such a study far enough to convince him that it is seldom if ever possible to posit simple polynomial regression solids in $(m - 1)$-way space $(m > 3)$, and finally obtain simple total regression curves—save, of course, in the classic all linear case. It seems altogether likely that we may ultimately be obliged to abandon polynomial forms of regression and use transcendental functions instead.

(*d*) If it indeed be true that simple regression surfaces for solids of three (or more) dimensions often imply very complicated total regression curves, it is a fact to be reckoned with in many applications of statistics. Consider, for example, the problem of business forecasting. One wishes to find the most likely value of a variable z for given values of variables x, y, etc. To fix the ideas, suppose z to be the money value of a certain industrial security six months hence, x the index of present demand for the article produced, and y an index of present[†] inflation.

Now suppose one has a considerable amount of data, and in the attempt to solve this problem first seeks the functional relationship between z and x only. Suppose it to be obvious from a glance at the graph thus obtained that it is very complicated. The practical man's judgment is that it is useless to try to fit it with a "mathematical curve," that a straight line is as good as anything, bad as that may be, and there is theoretical justification for that judgment. But he is apt to follow it up with the opinion that mathematics has no valuable place in this sort of problem anyhow, which is unfortunate. So far he has been dealing only with a total regression curve, z on x, but as a matter of fact z is really a function of two variables, x and y. It may well happen that the functional relationship between z and the two variables x and y is representable by a simple mathematical surface, and that this could be demonstrated if a solid correlation table were to be formed. I do not think that this method has been extensively tried. It involves a large amount of labour and of data, but it would seem to be a hopeful method of approach to the problem. In order to distinguish the two most important independent variables from among several it would be necessary to use multiple η, but this is not very much more difficult to compute than multiple r. I do not mean that methods of multiple correlation have not been tried, in the sense that the multiple correlation coefficient z on xy has been found, but this method we have already ruled out by supposing the total regressions non-linear. As we have seen, the only case where simple polynomial multiple regression leads to an equally simple total regression is the all linear case. It is peculiar in this respect. Therefore, lack of success in proceeding from simple linear to multiple linear regression should not discourage attempts to proceed from simple to multiple regression when the regressions are not linear.

---

[*] Seimatsu Narumi: "On the General Forms of Bivariate Frequency Distributions, etc.," *Biometrika*, Vol. xv. pp. 216, 217.

[†] In order that the author may make his meaning clear it is necessary to suppose a simple case, simpler than actually exists, as the reader will be well aware. Let it be understood also that these indexes of present states might if desired be cumulative indexes, obtained by observation not only of the present moment but also of the recent past.

MISCELLANEA.

On First Power Methods of finding Correlation.

By KARL PEARSON.

(1) *The "Linear" Ratio.* The measurement of association by the correlation ratio η suffers from the disadvantage that η, when we sample from material where there is no association, still shows a positive value; η^2 is the sum of a number of positive squares and thus its average value in a number of samples is not zero. If $\bar{y}_x$ be the mean of the array n_x and $\bar{y}$ the mean of the whole population, σ_y its standard deviation, and N its size, it would be a distinct advantage to have to deal with $S\left(\frac{n_x}{N}\frac{\bar{y}_x-\bar{y}}{\sigma_y}\right)$ instead of $S\left(\frac{n_x}{N}\left(\frac{\bar{y}_x-\bar{y}}{\sigma_y}\right)^2\right)$.

Now if we considered normal correlation and summed throughout the range, the above sum would be zero, but if we sum only for $\bar{y}_x > \bar{y}$ or for positive differences we have:

$$\frac{S^+ n_x(\bar{y}_x-\bar{y})}{N\sigma_y} = \frac{1}{\sqrt{2\pi}\,\sigma_x\sigma_y}\int_0^\infty e^{-\frac{1}{2}\frac{x^2}{\sigma_x^2}} r\frac{\sigma_y}{\sigma_x}\,x\,dx = \frac{1}{\sqrt{2\pi}}r \text{ ....(i).}$$

Similarly:

$$\frac{S^- n_x(\bar{y}_x-\bar{y})}{N\sigma_y} = -\frac{1}{\sqrt{2\pi}}r \text{ ....(ii).}$$

Or, we have:

$$r = \sqrt{\frac{\pi}{2}}\left(S\left(\frac{n_x}{N}\frac{\bar{y}_x-\bar{y}}{\sigma_y}\right)_{x>\bar{x}} + S\left(\frac{n_x}{N}\frac{\bar{y}-\bar{y}_x}{\sigma_y}\right)_{x<\bar{x}}\right) \text{ ....(iii).}$$

This is very simple theoretically, but it is not very effective in practice as we do not know in the case of *grouped* material* where the exact limit comes between $x>\bar{x}$ and $x<\bar{x}$, i.e. this point will probably be in the middle of an array. Let us call this array in which the mean, or for present purposes the median, occurs, the median array. Now in a portion of the median array $x>\bar{x}$ and in a second portion $x<\bar{x}$. What we have to determine are the contributions to the two sums of the parts of the median array on either side the median. Let the frequency of these parts be n_1 and n_2, the means $\bar{x}_1$, $\bar{y}_1$, $\bar{x}_2$, $\bar{y}_2$. On the marginal frequency for x let Nz_1/σ_x, Nz_0/σ_x and Nz_2/σ_x be the three ordinates of the normal curve, at the start, the median, and the finish of the median array. Then we require

$$\frac{n_1}{N}\frac{\bar{y}_1-\bar{y}}{\sigma_y} \text{ and } \frac{n_2}{N}\frac{\bar{y}_2-\bar{y}}{\sigma_y},$$

or using the regression line

$$r\frac{n_1}{N}\frac{\bar{x}_1-\bar{x}}{\sigma_x} \text{ and } r\frac{n_2}{N}\frac{\bar{x}_2-\bar{x}}{\sigma_x}.$$

But

$$\bar{x}_1-\bar{x} = \sigma_x\frac{(z_1-z_0)}{n_1/N} \text{ and } \bar{x}_2-\bar{x} = \sigma_x\frac{(z_0-z_2)}{n_2/N}.$$

Hence:

$$\frac{n_1}{N}\frac{\bar{y}_1-\bar{y}}{\sigma_y} = -r(z_0-z_1) \text{ and } \frac{n_2}{N}\frac{\bar{y}_2-\bar{y}}{\sigma_y} = r(z_0-z_2).$$

It follows therefore that

$$\begin{aligned} r\sqrt{\frac{2}{\pi}} = {} & \frac{Sn_x(y_x-\bar{y})}{N\sigma_y} \text{ above median array} \\ & \frac{Sn_x(\bar{y}-y_x)}{N\sigma_y} \text{ below median array} \\ & +r(2z_0-z_1-z_2) \text{ for median array.} \end{aligned}$$

*** For non-grouped material in which each individual is provided with an accurate measure of *x* and *y* this difficulty does not occur, but in such cases *r* can be found exactly by the product-moment method, and it is just in short series of data that it is desirable to use the most accurate method.**

But $z_0 = \frac{1}{\sqrt{2\pi}}$. Thus it follows that

$$r = \frac{1}{z_1 + z_2} \frac{S n_x (\bar{y}_x \sim \bar{y})}{N\sigma_y},$$ for all arrays, excluding median and taken always with the same sign, right and left of median* ....(iv).

The advantage of this formula is that except for the median array—which should be as small as feasible—it is independent of the number of arrays; for clearly any group of arrays can be clubbed together without modifying the value of r. There is accordingly no class index correction necessary. Further if r be really zero, there will be positive and negative values of $\bar{y}_x - \bar{y}$ on the right of median, and negative and positive values of $\bar{y} - \bar{y}_x$ on the left of median. Accordingly r in samples will in this case fluctuate about zero, being sometimes positive and sometimes negative; we have thus none of the difficulty which is associated with the correlation ratio, which can often take a quite sensible value—depending on the number of arrays where the true association is zero.

The disadvantages of the formula are that strictly it applies only to normal distributions of frequency, and therefore to linear regression and Gaussian variability. Practically, however, these limitations are less important than might be anticipated, and we will illustrate the method on cases which cannot be classed as normal in frequency.

Illustration I. The following values represent the mean place in class of 594 boys grouped in six intelligence classes:

	Very Able	Capable	Intelligent	Slow	Dull	Very Dull	Totals
Number	15	70	216	236	47	10	594
Mean	4·033	12·043	18·574	28·890	38·606	46·300	23·5875

The standard deviation of place in class $= 13·2501$.

Actually if all boys in the classes had been classified the mean place in class (reduced to classes of 50) should have been 25·0 and the standard deviation 14·4338. But the distribution instead of being rectangular tailed off considerably after class place 37, showing that the duller boys were not presented in due proportion for the intelligence estimation. We have still a very non-Gaussian distribution. Further the median category 216 is unfortunate, it is very large and z_0 is close to z_1. We found $z_1 = ·225,889$ and $z_2 = ·398,878$. Notwithstanding these disadvantages we found $r = ·6683$, while η_{PI} corrected for Class Index for intelligence was ·6780. This is very remarkable accordance considering the nature of the material.

To test the method further I took some cases where it was possible to apply directly the product-moment method, but it seems unnecessary to reproduce the tables in full. I will merely cite references.

Illustration II. 1112 Cases for Statures of Father and Son. See *Biometrika*, Vol. XVII. p. 219. From means of Son's stature for given stature of Father, there results: $r = ·5494$, where $\eta = ·5248$. From means of Father's stature for given stature of Son we have: $r = ·5552$, where $\eta = ·5268$. The difference of the two, ·0058, is practically of small importance. The actual value of the correlation found by the product-moment method is: $r = ·5206 \pm ·0147$. This shows the degree of divergence between the two processes for material, which forms as normal a bivariate distribution as we are likely to obtain in practice.

Illustration III. 2922 Cases of the contemporaneous Heights of the Barometer at Southampton and Laudale. We have here a case in which the distribution is very sensibly non-normal.

---

* Set as a problem in the London Honours Examination, B.Sc. Statistics, June, 1925.

The distribution will be found in the table on p. 291 of the present issue. Taking first the means of arrays of Laudale for constant Southampton we have: $r = \cdot 7927$, where $\eta = \cdot 7838$, and secondly from the means of arrays of Southampton for constant Laudale we find: $r = \cdot 7733$, where $\eta = \cdot 7808$. The mean of the two 'linear' ratios is ·7830. The actual value of the correlation as found by the product-moment method is $\cdot 7802 \pm \cdot 0050$.

It will be clear therefore that we do not get results substantially worse when the material is markedly skew than when the material is practically normal, notwithstanding the method—based on the normal curve—that we have applied to distribute the material in the median section.

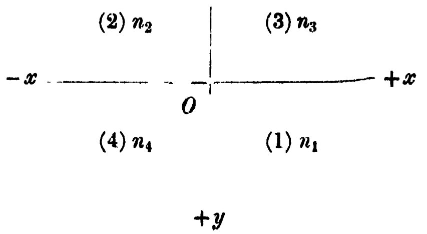
The method will, I think, be of value in cases where one variate is qualitative and the other quantitative. In this case the custom is to compute the correlation ratio; the present method, when the correlation ratio η is of order $\bar{\eta}$ (the mean value of η for no association), will throw some light on whether such η's are of significance or not; for our present coefficient should be zero if the value of η be merely the result of random sampling. It can serve in a limited way as a control on η. Its value will undoubtedly be increased, when the probable error of r found in this way has been determined. This I hope to provide shortly.

Instead of dividing our material into those individuals above and those below the mean we can distribute it into *four* classes instead of two, by reference to both means. This leads to an interesting series of formulae for the correlation, which I now proceed to deal with.

(2) "*Der deutsche Korrelations-Index*"—*and some others.* Let us find the centroid of the four volumes into which the normal surface

$$z = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)}$$

is divided by dichotomic planes through the x, z and y, z pairs of axes. We may call these quadrants, as marked in the diagram, (1), (2), (3) and (4), the corresponding frequencies n_1, n_2, n_3 and n_4, and the coordinates of the plans of the centroids $\bar{x}_1$, $\bar{y}_1$; $\bar{x}_2$, $\bar{y}_2$; $\bar{x}_3$, $\bar{y}_3$; and $\bar{x}_4$, $\bar{y}_4$ respectively. Clearly for a normal surface, we have $n_1 = n_2$, $\bar{x}_2 = -\bar{x}_1$, $\bar{y}_2 = -\bar{y}_1$ and $n_3 = n_4$, $\bar{x}_4 = -\bar{x}_3$, $\bar{y}_4 = -\bar{y}_3$. We may, however, with a view to later applications maintain the fuller nomenclatures.



Clearly:

$$n_1\bar{x}_1 = \int_0^\infty\int_0^\infty xz\,dx\,dy = \frac{N}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}\int_0^\infty\int_0^\infty xe^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)}dx\,dy,$$

$$n_1\bar{y}_1 = \int_0^\infty\int_0^\infty yz\,dx\,dy = \frac{N}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}\int_0^\infty\int_0^\infty ye^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)}dx\,dy.$$

Hence:

$$n_1\left(\frac{\bar{x}_1}{\sigma_1} - r\frac{\bar{y}_1}{\sigma_2}\right) = \frac{N}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}\int_0^\infty\int_0^\infty\left(\frac{x}{\sigma_1} - r\frac{y}{\sigma_2}\right)e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x}{\sigma_1} - r\frac{y}{\sigma_2}\right)^2}e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}}dx\,dy.$$

Let us now put $X=\frac{1}{\sqrt{1-r^2}}\left(\frac{x}{\sigma_1}-r\frac{y}{\sigma_2}\right)$ and integrate first with regard to X:

$$n_1\left(\frac{\bar{x}_1}{\sigma_1}-r\frac{\bar{y}_1}{\sigma_2}\right)=\int_0^\infty \frac{N\sqrt{1-r^2}}{2\pi\sigma_2}\left[-e^{-\frac{1}{2}X^2}\right]_0^{-\frac{r}{\sqrt{1-r^2}}\frac{y}{\sigma_2}} e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}}dy$$

$$=\int_0^\infty \frac{N\sqrt{1-r^2}}{2\pi\sigma_2}e^{-\frac{1}{2}\frac{1}{1-r^2}\frac{y^2}{\sigma_2^2}}dy=\frac{N(1-r^2)}{2\sqrt{2\pi}}.$$

Similarly:

$$n_1\left(\frac{\bar{y}_1}{\sigma_2}-r\frac{\bar{x}_1}{\sigma_1}\right)=\frac{N(1-r^2)}{2\sqrt{2\pi}}.$$

Or, solving, we find, what follows also from the symmetry (i.e. $\bar{x}_1/\sigma_1=\bar{y}_1/\sigma_2$):

$$n_1\frac{\bar{x}_1}{\sigma_1}=\frac{N(1+r)}{2\sqrt{2\pi}}=n_1\frac{\bar{y}_1}{\sigma_2} \quad \text{.......................................(v).}$$

Again:

$$n_3\bar{x}_3=\int_0^\infty\int_0^{-\infty}xz\,dx=\frac{N}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}\int_0^\infty\int_0^{-\infty}xe^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}-\frac{2rxy}{\sigma_1\sigma_2}+\frac{y^2}{\sigma_2^2}\right)}dx\,dy$$

$$-\frac{1}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}\int_0^\infty\int_0^\infty xe^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}-2(-r)\frac{xy}{\sigma_1\sigma_2}+\frac{y^2}{\sigma_2^2}\right)}dx\,dy$$

by changing y to $-y$. Thus we reach the same result as before by changing r to $-r$; or

$$n_3\frac{\bar{x}_3}{\sigma_1}=\frac{N(1-r)}{2\sqrt{2\pi}}=-n_3\frac{\bar{y}_3}{\sigma_2} \quad \text{.......................................(vi).}$$

Now let x_s, y_s be observations falling in the sth quadrant. Then

$$\left.\begin{array}{ll}
S(x_1)=n_1\bar{x}_1=\frac{\sigma_1N(1+r)}{2\sqrt{2\pi}}, & S(x_2)=n_2\bar{x}_2=-\frac{\sigma_1N(1+r)}{2\sqrt{2\pi}} \\
S(x_3)=n_3\bar{x}_3=\frac{\sigma_1N(1-r)}{2\sqrt{2\pi}}, & S(x_4)=n_4\bar{x}_4=-\frac{\sigma_1N(1-r)}{2\sqrt{2\pi}} \\
S(y_1)=n_1\bar{y}_1=\frac{\sigma_2N(1+r)}{2\sqrt{2\pi}}, & S(y_2)=n_2\bar{y}_2=-\frac{\sigma_2N(1+r)}{2\sqrt{2\pi}} \\
S(y_3)=n_3\bar{y}_3=-\frac{\sigma_2N(1-r)}{2\sqrt{2\pi}}, & S(y_4)=n_4\bar{y}_4=\frac{\sigma_2N(1-r)}{2\sqrt{2\pi}}
\end{array}\right\} \quad \text{.........(vi } bis\text{).}$$

These equations by elimination of σ_1 and σ_2 give us a variety of methods for finding r on the assumption of normality.

For example:

$$r=\frac{S(x_1)-S(x_3)}{S(x_1)+S(x_3)}=\frac{S(x_2)-S(x_4)}{S(x_2)+S(x_4)} \quad \text{.........................(vii } a \text{ \& } b\text{),}$$

$$r=\frac{S(y_1)-S(y_4)}{S(y_1)+S(y_4)}=\frac{S(y_2)-S(y_3)}{S(y_2)+S(y_3)} \quad \text{.........................(viii } a \text{ \& } b\text{),}$$

whence:

$$r=\frac{S(x_1)-S(x_2)-S(x_3)+S(x_4)}{S(x_1)-S(x_2)+S(x_3)-S(x_4)} \quad \text{...............................(ix } a\text{),}$$

$$=\frac{S(y_1)-S(y_2)+S(y_3)-S(y_4)}{S(y_1)-S(y_2)-S(y_3)+S(y_4)} \quad \text{...............................(ix } b\text{).}$$

Or, again we may take:

$$r=\frac{S(x_1)-S(x_2)+S(x_4)+S(y_3)-S(x_3)+S(y_1)-S(y_2)-S(y_4)}{S(x_1)-S(x_2)+S(x_3)-S(x_4)+S(y_1)-S(y_2)-S(y_3)+S(y_4)} \quad \text{...............(x),}$$

or still again:

$$r=\tfrac{1}{2}\left\{\frac{S(x_1)-S(x_3)}{S(x_1)+S(x_3)}+\frac{S(x_2)-S(x_4)}{S(x_2)+S(x_4)}\right\} \quad\ldots\ldots\ldots\ldots\ldots\ldots\text{(xi)},$$

$$=\tfrac{1}{2}\left\{\frac{S(y_1)-S(y_4)}{S(y_1)+S(y_4)}+\frac{S(y_2)-S(y_3)}{S(y_2)+S(y_3)}\right\} \quad\ldots\ldots\ldots\ldots\ldots\ldots\text{(xii)},$$

$$=\tfrac{1}{2}\left\{\frac{S(x_1)-S(x_2)-S(x_3)+S(x_4)}{S(x_1)-S(x_2)+S(x_3)-S(x_4)}+\frac{S(y_1)-S(y_2)+S(y_3)-S(y_4)}{S(y_1)-S(y_2)-S(y_3)+S(y_4)}\right\} \quad\ldots\ldots\text{(xiii)}.$$

The right-hand side of the last result has been termed by Lenz "Der deutsche Korrelations-Index*." He appears to think that it has some special merits which he does not, however, reveal, beyond stating that it is easier to compute than the usual correlation coefficient. As, however, he appears to compute the usual correlation coefficient by subtracting each deviation from its mean, there is no wonder that he has found the process laborious†. Lenz's "deutsche Korrelations-Index" is the usual correlation coefficient when the distribution is normal, or approximately normal, a result he does not seem to have been aware of. It is really only valid for such cases. What he then has reached is a German method of *finding* the usual correlation coefficient, and whether it is better or worse than any other method will depend on its probable error. This he has not determined. It is true he does give a probable error for each value cited of the "deutsche Korrelations-Index" but it is calculated from the formula for the probable error of the product-moment coefficient, and therefore is entirely fallacious. The probable error as in Sheppard's formula, which is far simpler of application than Lenz's, will probably be found to average 30 % to 50 % more than that of the product-moment method, which has the minimum probable error of all methods discoverable. Lenz asks why sums of squares and products should be used when it is so much easier to calculate sums of deviations‡. The answer is an easy one, it is done because the method is more accurate.

The advantage of the correlation coefficient lies—whatever be the distribution—in its geometrical properties, i.e. its relation to the slope of the best-fitting straight line and the expression by aid of this coefficient of the mean square deviation from this line. The "deutsche Korrelations-Index" is simply this coefficient if the distribution be practically normal. If the distribution be not normal, nobody can say what the coefficient means; it does not as far as I can see express any fundamental geometrical properties of the frequency distribution.

---

* Lenz gives no proof whatever of his formula, he simply writes it down (*Archiv für Hygiene*, Bd. XCIII. S. 147) after a criticism of the product-moment correlation coefficient, which is wholly fallacious.

† The first correlation coefficients calculated by the product-moment method were those given in my third memoir on evolution of 1896; the product-moments as well as the standard deviations were on that occasion calculated, as every trained mathematician would calculate them, about an arbitrary origin and then transferred to the mean. It is therefore somewhat amusing to read in a recent American text-book that this great discovery of reduction to the mean was made by an American in 1917! Every mathematician and engineer has used such reduction since "radii of gyration" and "principal axes" were discovered in the eighteenth century.

‡ I do not propose to criticise Lenz's sociological hypotheses, but I would criticise his statistical methods. He has eleven correlation tables. Of these tables one has 54 and one 20 entries, the remainder have 8, 10, 9, 8, 10, 10, 8, 10 and 14 entries respectively. To all these he applies the probable error of the correlation coefficient on the assumption that he is dealing with a large sample. It may be safely said that it is *totally* inapplicable to all the numbers dealt with, excepting possibly the first two, but here the intensity of the correlation is such, −·86 for the 54 (Bayern) and −·72 for the 20 (Deutschland), that the probable error as found for *large* samples is without meaning. The only case with sufficient pairs to give a reasonable result for the correlation is that of Bayern, and here the German correlation coefficient is close to the product-moment value.

Returning to our Equations (vi *bis*) we note that

$$\left.\begin{aligned} S(x_1)-S(x_3) &= \frac{\sigma_1 N r}{\sqrt{2\pi}} = \quad S(x_4)-S(x_2) \\ S(y_1)-S(y_4) &= \frac{\sigma_2 N r}{\sqrt{2\pi}} = \quad S(y_3)-S(y_2) \\ S(x_1)+S(x_3) &= \frac{\sigma_1 N}{\sqrt{2\pi}} = -\{S(x_2)+S(x_4)\} \\ S(y_1)+S(y_4) &= \frac{\sigma_2 N}{\sqrt{2\pi}} = -\{S(y_2)+S(y_3)\} \end{aligned}\right\} \text{......................(xiv).}$$

Accordingly:

$$\frac{S(x_1)-S(x_3)}{S(y_1)+S(y_4)} = r\frac{\sigma_1}{\sigma_2} = \frac{S(x_2)-S(x_4)}{S(y_2)+S(y_3)} \text{ ....(xv),}$$

$$\frac{S(y_1)-S(y_4)}{S(x_1)+S(x_3)} = r\frac{\sigma_2}{\sigma_1} = \frac{S(y_2)-S(y_3)}{S(x_2)+S(x_4)} \text{ ....(xvi).}$$

We have thus two expressions from each of which, or from the mean of each pair, the two regression coefficients can be found. These expressions give us a series of values again for the correlation coefficient such as:

$$r^2 = \frac{S(x_1)-S(x_3)}{S(x_1)+S(x_3)} \cdot \frac{S(y_1)-S(y_4)}{S(y_1)+S(y_4)} = \frac{S(x_2)-S(x_4)}{S(x_2)+S(x_4)} \cdot \frac{S(y_2)-S(y_3)}{S(y_2)+S(y_3)} \text{ ...(xvii } a \text{ \& } b\text{),}$$

or, r may be taken as the fourth-root of all four products. But again:

$$\frac{\sigma_1}{\sigma_2} = \frac{S(x_1)}{S(y_1)} = \frac{S(x_2)}{S(y_2)} = -\frac{S(x_3)}{S(y_3)} = -\frac{S(x_4)}{S(y_4)} \text{ ....(xviii).}$$

Hence:

$$r = \frac{1-\dfrac{S(x_3)}{S(x_1)}}{1+\dfrac{S(y_4)}{S(y_1)}} = \frac{1-\dfrac{S(x_4)}{S(x_2)}}{1+\dfrac{S(y_3)}{S(y_2)}} = \text{etc. ....(xix } a\text{),}$$

$$= \frac{1-\dfrac{S(y_4)}{S(y_1)}}{1+\dfrac{S(x_3)}{S(x_1)}} = \frac{1-\dfrac{S(y_3)}{S(y_2)}}{1+\dfrac{S(x_4)}{S(x_2)}} = \text{etc. ....(xix } b\text{).}$$

Further we have:

$$\sigma_1\sqrt{1-r^2} = 2\sqrt{2\pi}\sqrt{\frac{S(x_1)}{N}\cdot\frac{S(x_3)}{N}} = 2\sqrt{2\pi}\sqrt{\frac{S(x_2)}{N}\cdot\frac{S(x_4)}{N}} \text{ ....(xx } a\text{),}$$

$$\sigma_2\sqrt{1-r^2} = 2\sqrt{2\pi}\sqrt{\frac{S(y_1)}{N}\cdot\frac{S(y_4)}{N}} = 2\sqrt{2\pi}\sqrt{\frac{S(y_2)}{N}\cdot\frac{S(y_3)}{N}} \text{ ....(xx } b\text{).}$$

It is thus possible to find the mean square variability of the arrays about the regression line. Since r is known, we can also find σ_1 and σ_2.

The fact that from the eight sums we have eight equations to find three quantities, σ_1, σ_2 and r, shows that we have ample evidence to measure the degree of approach to a normal distribution. For we ought to have

$$S(x_1) = -S(x_2), \quad S(x_3) = -S(x_4), \quad S(y_1) = -S(y_2), \quad S(y_3) = -S(y_4),$$

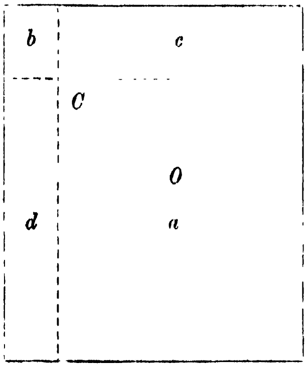
but these equations will be rarely accurately fulfilled, especially in the case of small samples. It is therefore advisable to average out the different values obtained for the constants.

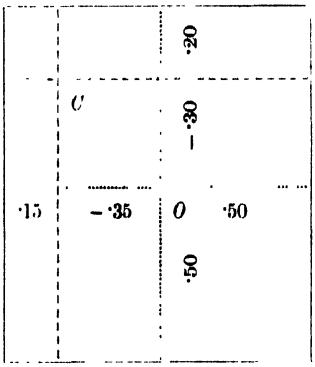
A difficulty, however, arises as in our previous discussion. It is only possible to deal with individual measurements in the case of relatively small samples*. Most observations ar grouped, and there arises the difficulty of distributing the individuals in the cell (often one of large content) in which the mean lies, and in the arrays through which the mean dichotomic planes pass. If we assume a normal surface and endeavour to integrate out the volumes in question, we obtain long series in r and the tetrachoric functions, so that the final value of r can only be ascertained by approximation. It is worth while determining what sort of accuracy we get by computing the correlation by these formulae on a long series, which diverges not too much from normality and where the product-moment coefficient is known. I will take the table for correlation of stature in Father and Son provided by Professor Yasukawa on p. 218 of the current issue*. I have surrounded with heavy lines the column and row in which the mean statures of Father and Son lie: see Table, p. 466. It is clear that to obtain our summations, i.e. $S(x)$ and $S(y)$ for the quadrants, we must proportion the central column, the central row, and their intersecting cell. As I have said, the true proportioning depends on a knowledge of r itself and, even supposing we take an approximate value of r, is extraordinarily laborious. I shall therefore content myself with assuming that the contents of the cells are rectanguloid blocks and taking as proportions the volumes standing upon the corresponding areas. Even so, the total labour of proportioning and determining the means of the quadrants is three or four times as great as that of finding the product-moment correlation, and does not provide the other requisite constants of the distribution.

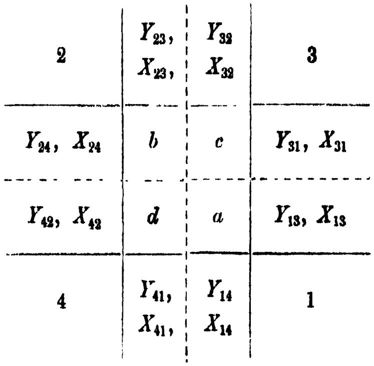
---

* Since these methods involve knowing the means and the actual measurements of each individual, we can classify at once into four groups and use Sheppard's result, which does not involve adding measurements but only counting the *number* of individuals. But here again product-moment methods would be of higher accuracy, and ultimately (with grouping) shorter for long series.

Now the central line of the emphasised column is at 5′ 9″ stature of Son, or the mean (5′ 8″, ·65) is at − ·35 of our unit, an inch, from this central line. Similarly the mean stature of Father is − ·30 from the central line of the emphasised row. Accordingly the coordinates of the mean referred to our working origin O are − ·35 and − ·30. We have first to proportion the contents, 25·75, of the central cell into four groups, a, b, c, d. The areas of the bases of a, b, c, d are as ·80 × ·85, ·15 × ·20, ·85 × ·20, and ·80 × ·15, or as ·6800, ·0300, ·1700 and ·1200 respectively. The







frequencies will be 17·5100, ·7725, 4·3775, 3·0900. The coordinates referred to C will be + ·425, + ·400 for a, − ·075, − ·10 for b, + ·425, − ·10 for c and − ·075, + ·400 for d. These provide the following contributions to the first moments of the four quadrants about the axes through C:

$$\begin{aligned} a_x &= +7{\cdot}441{,}7500, & a_y &= +7{\cdot}004{,}00, \\ b_x &= -\ {\cdot}057{,}9375, & b_y &= -\ {\cdot}077{,}25, \\ c_x &= +1{\cdot}860{,}4375, & c_y &= -\ {\cdot}437{,}75, \\ d_x &= -\ {\cdot}231{,}7500, & d_y &= +1{\cdot}236{,}00. \end{aligned}$$

We have next to obtain with the same form of proportioning the contributions to the x and y moments of the emphasised column, now *omitting* the central cell.

---

* Our results are corrected for grouping. For Professor Yasukawa's purposes those corrections were not used.

TABLE.

Correlation between Father's Stature and Son's Stature.

Son's Stature = x.

Father's Stature = y.

	4'11"½—5'0"½	5'0"½—5'1"½	5'1"½—5'2"½	5'2"½—5'3"½	5'3"½—5'4"½	5'4"½—5'5"½	5'5"½—5'6"½	5'6"½—5'7"½	5'7"½—5'8"½	5'8"½—5'9"½	5'9"½—5'10"½	5'10"½—5'11"½	5'11"½—6'0"½	6'0"½—6'1"½	6'1"½—6'2"½	6'2"½—6'3"½	6'3"½—6'4"½	6'4"½—6'5"½	6'5"½—6'6"½	6'6"½—6'7"½	6'7"½—6'8"½	Totals
4'10"½—4'11"½	—	—	—	—	1	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3·00
4'11"½—5'0"½	—	—	·25	·25	—	1	·5	1·5	—	—	—	—	—	—	—	—	—	—	—	—	—	3·50
5'0"½—5'1"½	—	—	·25	·25	1·5	·5	1	1·5	1	1	—	—	—	—	—	—	—	—	—	—	—	7·00
5'1"½—5'2"½	—	—	·5	2·75	3·75	2	2·25	5·25	2·5	—	—	—	—	—	—	—	—	—	—	—	—	19·00
5'2"½—5'3"½	·5	·5	·5	2·25	4	2·75	6·25	4·5	7·75	5	1	—	—	—	1	—	—	—	—	—	—	36·00
5'3"½—5'4"½	·5	—	2	2	4·25	7·5	11·5	12·75	11	5·5	3	3·25	·5	·25	—	—	—	—	—	—	—	64·00
5'4"½—5'5"½	1	·25	·25	3·75	7·75	12·5	11	20·75	10	11·75	6·25	5	3	·75	1·5	—	—	—	—	—	—	95·50
5'5"½—5'6"½	—	·25	·25	4·75	8·5	10·75	17·5	27·25	25	19·75	18·75	7·75	1	·5	1·5	—	—	—	—	—	—	143·50
5'6"½—5'7"½	—	1	·5	2·75	3·5	10·5	16·75	26·5	32·75	17·25	14·75	11·25	8	2·5	—	—	—	—	—	—	—	148·00
5'7"½—5'8"½	—	—	·5	1·25	2·25	5·5	15·5	20·5	25	25·75	20	17·25	9·75	6·5	6·25	2	1·25	·25	—	—	—	159·50
5'8"½—5'9"½	—	—	—	—	2·5	2·5	5·25	14	28·75	29	21	16·5	11·25	7	2·75	2	·25	·25	—	—	—	143·00
5'9"½—5'10"½	—	—	—	·25	·75	2·5	2	15·25	12·5	21·75	20	20·25	12·75	6·5	2·5	—	—	—	1	·5	·5	119·00
5'10"½—5'11"½	—	—	—	·25	1·25	—	2·5	3·25	9·5	10·5	13·5	11·25	10·5	8	6·5	2·5	·5	—	—	—	—	80·00
5'11"½—6'0"½	—	—	—	—	—	—	1	1	9	4·5	6·25	9	8·5	7·25	3·25	·75	1·5	·5	·25	·25	—	53·00
6'0"½—6'1"½	—	—	—	—	—	—	—	1	2·25	2·25	3·5	5	3·75	3·25	2·75	2·25	1	·5	·75	·25	—	28·50
6'1"½—6'2"½	—	—	—	—	—	—	—	—	—	—	1·5	1	·5	·5	—	·5	—	—	—	—	—	4·00
6'2"½—6'3"½	—	—	—	—	—	—	—	—	—	1	1	1	—	2·5	—	—	—	—	—	—	—	5·50
Totals	2·00	2·00	5·00	20·50	41·00	60·00	93·00	155·00	177·00	155·00	130·50	108·50	69·50	45·50	28·00	10·00	4·50	1·50	2·00	1·00	·50	1112

$m_x = 5'8''\cdot6511$ $\sigma_x = 2''\cdot7289$ $r_{xy} = \cdot5206 \pm \cdot0147$

$m_y = 5'7''\cdot6965$ $\sigma_y = 2''\cdot7018$

The notation for these contributions is given in the third figure, the quadrants being numbered 2, 3, 1, 4. Thus Y_{23} signifies the moment round the centroid axis of x, which is contributed to quadrant 2 by the column between quadrants 2 and 3, X_{31} is the moment round the centroid axis of y which is contributed to quadrant 3 by the row which lies between quadrants 3 and 1. And so on. Clearly X_{23}, X_{32}, X_{14} and X_{41} are very simply found. We have only to proportion the total of the parts of the central column in the ratio ·15 to ·85, and multiply by the arms $-{\cdot}075$ and $+{\cdot}425$ as the case may be. Again Y_{24}, Y_{31}, Y_{42} and Y_{13} admit of being found in like manner. In the case of Y_{23}, Y_{32}, Y_{41}, Y_{14}, X_{24}, X_{31}, X_{13} and X_{42}, it is best to compute the moments of the portions of the rows and columns first about the working axes, and then reduce these to the centroid, and finally proportion them in the ratio of the given parts they are of the total column or row. Thus consider Y_{23}. We take the frequencies of upper portion of central column:

$$\begin{array}{lcl}
1 \quad \times -7 = -\ \ 7 & & \\
- \quad \times -6 = \ \ - & & -146{\cdot}0 = \text{value about } O, \\
5 \quad \times -5 = -\ 25 & & \\
5{\cdot}5 \ \times -4 = -\ 22 & & -146{\cdot}0 + {\cdot}30 \times 60{\cdot}25 = -127{\cdot}925, \\
11{\cdot}75 \times -3 = -\ 35{\cdot}25 & & \\
19{\cdot}75 \times -2 = -\ 39{\cdot}50 & & = \text{value about } C. \\
17{\cdot}25 \times -1 = -\ 17{\cdot}25 & & \\
\hline
60{\cdot}25 \qquad\quad -146{\cdot}0 & &
\end{array}$$

Proportioning in the ratios of ·15 and ·85, we find

$$Y_{23} = -19{\cdot}18875, \qquad Y_{32} = -108{\cdot}73625.$$

In this manner all the column and row contributions were found, straightforwardly, but laboriously. We have:

$$\begin{array}{ll}
Y_{23} = -19{\cdot}18875, & X_{23} = -{\cdot}677{,}8125, \\
Y_{32} = -108{\cdot}73625, & X_{32} = +21{\cdot}765{,}3125, \\
Y_{31} = -1{\cdot}26500, & X_{31} = +37{\cdot}177{,}500, \\
Y_{13} = +20{\cdot}24000, & X_{13} = +148{\cdot}710{,}000, \\
Y_{14} = +24{\cdot}1425, & X_{14} = +24{\cdot}926{,}250, \\
Y_{41} = +136{\cdot}8075, & X_{41} = -{\cdot}776{,}250, \\
Y_{42} = -1{\cdot}41000, & X_{42} = -105{\cdot}660{,}000, \\
Y_{24} = +22{\cdot}56000, & X_{24} = -26{\cdot}415{,}000.
\end{array}$$

The next stage was to determine the moments of the quadrants apart from the proportioned central column, central row and central cell. These were found about the working origin, and then reduced to the mean. We will call them $S_t(x)$, $S_t(y)$, where t refers to the quadrants. I found:

$$\begin{array}{ll}
S_1(x) = +765{\cdot}8625, & S_1(y) = +741{\cdot}775, \\
S_2(x) = -908{\cdot}7875, & S_2(y) = -936{\cdot}175, \\
S_3(x) = +203{\cdot}7750, & S_3(y) = -155{\cdot}550, \\
S_4(x) = -167{\cdot}7125, & S_4(y) = +265{\cdot}175.
\end{array}$$

We are now prepared to combine our results so as to obtain the first moments of each quadrant about both axes. The formulae are:

$$\begin{aligned}
S(x_1) &= S_1(x) + X_{13} + X_{14} + a_x, \\
S(x_2) &= S_2(x) + X_{23} + X_{24} + b_x, \\
S(x_3) &= S_3(x) + X_{32} + X_{31} + c_x, \\
S(x_4) &= S_4(x) + X_{42} + X_{41} + d_x, \\
S(y_1) &= S_1(y) + Y_{13} + Y_{14} + a_y, \\
S(y_2) &= S_2(y) + Y_{23} + Y_{24} + b_y, \\
S(y_3) &= S_3(y) + Y_{32} + Y_{31} + c_y, \\
S(y_4) &= S_4(y) + Y_{42} + Y_{41} + d_y.
\end{aligned}$$

The reader who has examined the magnitudes of the proportioned terms given above will see how impossible it is to neglect the central row, column or cell*. Their contributions cannot be disregarded, and, even if our rough approximations are allowable, they make the work far more laborious for grouped tables—the usual condition of material—than directly finding the product-moment value.

We deduce:

$S(x_1) = +946{\cdot}94050, \quad S(y_1) = +905{\cdot}82650,$

$S(x_2) = -935{\cdot}93825, \quad S(y_2) = -956{\cdot}85100,$

$S(x_3) = +264{\cdot}57825, \quad S(y_3) = -265{\cdot}98900,$

$S(x_4) = -274{\cdot}38050, \quad S(y_4) = +313{\cdot}11350.$

These give a position of the true mean at $-{\cdot}349$ and $-{\cdot}304$, i.e. they do not add up to exactly zero, because we only used the mean statures to hundredths and not thousandths of an inch.

We can now determine the values as found by the formulae of the preceding paper.

First, the correlation coefficient r. We have:

$r = {\cdot}5632$ (from vii a), $\quad r = {\cdot}5466$ (from vii b),
$r = {\cdot}4863$ (from viii a), $\quad r = {\cdot}5650$ (from viii b),
$r = {\cdot}5549$ (from ix a), $\quad r = {\cdot}5257$ (from ix b),
$r = {\cdot}5402$ (from x),
$r = {\cdot}5549$ (from xi), $\quad r = {\cdot}5256$ (from xii),
$r =$ "Der deutsche Korrelations-Index" $= {\cdot}5403$ (from xiii),
$r = {\cdot}5233$ (from xvii a), $\quad r = {\cdot}5557$ (from xvii b),
$r = {\cdot}5355 = {\cdot}5114 =$ etc. (from xix a),
$r = {\cdot}5531 = {\cdot}5583 =$ etc. (from xix b).

Here are 13 values at least of r[†], determined from linear moments. Which does the reader prefer? The true correlation coefficient as given by the product-moment formula, i.e. ·5206 ± ·0147, is *unique*; it depends on no special form of distribution, and has a perfectly clear geometrical meaning. It gives no choice to the user. Here we have a great number of variants—in the bulk significantly higher than the coefficient of correlation. They have no geometrical interpretation except as approximations to the correlation coefficient, and this, as far as we yet know, only for the case of normal frequency. At present no probable errors have been determined, and all we can at present say is, that if we assume normal correlation, they will be higher than that of the correlation coefficient found by product-moment. The work of deducing one or other of them is considerably more laborious and yet less exact than that of the product-moment method[‡], for our proportioning is only rough and we do not know, as in the latter case, what (if any) corrections should be made for "grouping." No wise computer would think of using any of these linear moment methods for a long series, in which both variates were measurable characters. On the other hand, they may be serviceable as controls on other processes when one variate is qualitative, as shown in the example on p. 460. If there be only a few observations —taken to a high degree of accuracy—then these methods are easy of application; but it is precisely in such cases that we ought to apply the product-moment method, for it gives us the least probable error for the given size of the sample. What is more, it provides in the course of the work additional constants, which are very generally needed.

---

[*] If we neglect them we find from (vii a & b) and (viii a & b) ·5797, ·6884, ·5254 and ·7150 respectively! This shows how important proportioning is, and better proportioning might possibly reduce the results still closer to the true value.

[†] It is not, perhaps, legitimate to consider (ix a) and (xi) or (ix b) and (xii) as giving independent values, and even some of the others can hardly be looked upon as more than averaging processes.

[‡] I assume that the computer, of course, knows how to apply the product-moment correctly. He might spend any amount of time over 1115 observations, if he subtracted the characters of each from their means!

It may be said that this additional information would be provided by our Equations (xv), (xvi), (xviii) and (xx *a* & *b*). Let us examine this. The unique values of the product-moment regression coefficients are ·5258 and ·5104, and the standard deviations are $\sigma_1 = 2{\cdot}7289$ and $\sigma_2 = 2{\cdot}7018$.

From (xv) $\quad r\sigma_1/\sigma_2 = {\cdot}5598 = {\cdot}5410.$

From (xvi) $\quad r\sigma_2/\sigma_1 = {\cdot}4892 = {\cdot}5708.$

These results are *inter se* discordant, and their means ·5504, ·5300 are considerably in excess of the true values. The ratio σ_1/σ_2 has for its true value 1·0100. From (xviii), we have for its values:

$$\frac{\sigma_1}{\sigma_2} = 1{\cdot}0325 = {\cdot}97075 = 1{\cdot}3100 = {\cdot}63246,$$

a very discordant series, the mean of which, ·9864, would be interpreted as Sons being less variable than Fathers—a result contrary to what has been observed.

Lastly, before applying (xx *a* & *b*), we must adopt some value of r from our discordant returns. Let us take the value ·5403 provided by the "deutsche Korrelations-Index." We find from (xx *a*):

$$\sigma_1 = 2{\cdot}682 = 2{\cdot}715. \qquad \text{(True value } 2{\cdot}7289.)$$

$$\sigma_2 = 2{\cdot}853 = 2{\cdot}703. \qquad \text{(True value } 2{\cdot}7018.)$$

One value of σ_1 is less than one of σ_2, and another value of σ_1 is greater than one of σ_2, and the mean values $\sigma_1 = 2{\cdot}698$ and $\sigma_2 = 2{\cdot}778$ are far from acceptable. As in any particular example we do not know which are the better of the alternatives, we must either guess at random or take means. It will thus be seen that although other constants, as well as r, can be determined by linear moment methods, these lead to divergent results, and results not in agreement with the true values. It is conceivable that with some other system of proportionment, more self-accordant results might be obtained, and these more in agreement with direct method values. But we doubt this, at least for any case which has not an exact normal distribution. And even if such a method of proportioning could be found, we believe the labour of computing would be still more considerable. It is hard to believe that even a reasonable national pride will be able to justify the use of "der deutsche Korrelations-Index." As far as I am aware, nobody has given the product-moment coefficient of correlation a national name, and I trust they will not. National feeling has little to do with true science.

Further consideration of the Integral $\int_0^\theta \cos^{n+1}\theta\, d\theta$ for large values of n.

By JOHN WISHART, M.A., B.Sc.

THE method of my earlier paper* works very well for the range of n illustrated therein; i.e. for n above 100. For lower values of n the approximation is not so good. That, however, in the long run, does not matter, as the integral in question, for any value of n below 100, will ultimately be given by the *Tables of the Incomplete Beta-Function* now in preparation. But the calculation of these Tables for the full range of the arguments is naturally taking some little time. In the meantime recourse may be had to another approximate expansion for the integral, which has been discovered since the above paper was written. It has been found to give excellent results, over the full range of 3σ from the mode, for n as low as 8. For such low values it is, however, shorter to expand by the binomial theorem, or to use one of the well-known expansions given in trigonometrical text-books.

---

* *Biometrika*, Vol. XVI. pp. 68—78.

Briefly the method is as follows:

Replace $n+1$ by N, and adopt the transformation $t=\tan\frac{\theta}{2}$. Then

$$\int_0^\theta \cos^N\theta\, d\theta = 2\int_0^t \left(\frac{1-t^2}{1+t^2}\right)^N \frac{dt}{1+t^2} \qquad 0 \leqslant t \leqslant 1 \text{ ....(1)},$$

$$= \int_0^t \frac{1}{1+t^2} e^{-2N(t^2+\frac{1}{3}t^6+\frac{1}{5}t^{10}+\dots)}\,dt.$$

Let $t=\frac{x}{2\sqrt{N}}$. Then

$$\int_0^\theta \cos^N\theta\, d\theta = \frac{1}{\sqrt{N}} \int_0^x \frac{1}{1+\frac{x^2}{4N}} e^{-\frac{x^2}{2}-\frac{x^6}{96N^2}-\frac{x^{10}}{2560N^4}-\dots}\,dx,$$

$$= \frac{1}{\sqrt{N}} \int_0^x e^{-\frac{1}{2}x^2} \times \left(1-\frac{x^2}{4N}+\frac{x^4}{16N^2}-\dots\right)\left(1-\frac{x^6}{96N^2}-\frac{x^{10}}{2560N^4}+\frac{x^{12}}{18432N^4}+\dots\right)dx,$$

$$= \frac{1}{\sqrt{N}} \int_0^x e^{-\frac{1}{2}x^2} \times \left[1-\frac{x^2}{4N}+\frac{x^4}{16N^2}-\frac{x^6}{32N^2}\left(\frac{1}{3}+\frac{1}{2N}\right)+\frac{x^8}{128N^3}\left(\frac{1}{3}+\frac{1}{2N}\right) - \frac{x^{10}}{960N^4}+\frac{x^{12}}{18432N^4}\right] \text{...(2)},$$

neglecting terms of higher order than $\frac{1}{N^4}$.

Expressing in terms of the Incomplete Normal Moment Functions as before, we have

$$I_\theta(N) = \sqrt{\frac{N}{2}} \cdot \frac{\Gamma(\frac{1}{2}N)}{\Gamma\{\frac{1}{2}(N+1)\}} \times \left[m_0(x) - \frac{1}{4N}m_2(x) + \frac{3}{16N^2}m_4(x) - \frac{15}{32N^2}\left(\frac{1}{3}+\frac{1}{2N}\right)m_6(x) + \frac{105}{128N^3}\left(\frac{1}{3}+\frac{1}{2N}\right)m_8(x) - \frac{63}{64N^4}m_{10}(x) + \frac{1155}{2048N^4}m_{12}(x)\right] \text{...(3)},$$

where $I_\theta(N)$ is the probability integral of the symmetrical frequency curves, as defined in the previous paper.

Hence

$$I_\theta(N) = \sqrt{\frac{N}{2}} \cdot \frac{\Gamma(\frac{1}{2}N)}{\Gamma\{\frac{1}{2}(N+1)\}} \times \left[\phi_0(x) - \frac{1}{N}\phi_1(x) + \frac{1}{N^2}\phi_2(x) - \frac{1}{N^3}\phi_3(x) + \frac{1}{N^4}\phi_4(x)\right] \text{...(4)},$$

where

$\phi_0(x) = m_0(x) = \frac{1}{2}\alpha_x$,

$\phi_1(x) = {\cdot}25m_2(x)$,

$\phi_2(x) = {\cdot}1875m_4(x) - {\cdot}15625m_6(x)$,

$\phi_3(x) = {\cdot}234375m_6(x) - {\cdot}2734375m_8(x)$,

$\phi_4(x) = {\cdot}41015625m_8(x) - {\cdot}984375m_{10}(x) + {\cdot}56396484m_{12}(x)$,

and

$$x = 2\sqrt{N}\tan\frac{\theta}{2} = \frac{2\sqrt{N}}{\sin\theta}(1-\cos\theta).$$

A Table of these functions has been calculated from the existing Table of Moment Functions*, and appears on p. 471.

It may be noted that $\sqrt{\frac{N}{2}} \cdot \frac{\Gamma(\frac{1}{2}N)}{\Gamma\{\frac{1}{2}(N+1)\}} = \sqrt{\frac{n}{n+1}} \cdot c_0$, where c_0 is the first coefficient of the earlier expansion, and is tabled from $n=100$ to $n=400$. An alternative method of evaluating this expression is to use the *Tables of the Gamma-Function*†. Or again we have

$$\sqrt{\frac{N}{2}} \cdot \frac{\Gamma(\frac{1}{2}N)}{\Gamma\{\frac{1}{2}(N+1)\}} = \sqrt{\frac{\pi}{2N}} \Big/ \int_0^{\frac{\pi}{2}} \cos^N\theta\, d\theta,$$

the complete integral being tabled from $N=0$ to $N=104$ in *Biometrika*, Vol. XI. p. 377.

---

* *Tables for Statisticians and Biometricians*, Table IX.

† *Tracts for Computers*, No. VIII.

Table of the Functions $\phi(x)$ *for computing* $\int_0^\theta \cos^N \theta \, d\theta$, *where* $x = 2\sqrt{N} \tan \frac{1}{2}\theta$.

x	$\phi_0(x)$	δ^2	δ^4	$\phi_1(x)$	δ^2	$\phi_2(x)$	δ^2	$\phi_3(x)$	δ^2	$\phi_4(x)$	δ^2
0·0	·00000000	0	0	·0000000	0	·0000000	0	·0000000	0	·0000000	0
0·1	·03982784	− 39597	+1178	·0000331	+1966	·0000000	+ 16	·0000000	0	·0000000	0
0·2	·07925971	− 78016	+2296	·0002628	+3813	·0000016	+ 84	·0000000	+ 2	·0000000	0
0·3	·11791142	−114139	+3303	·0008738	+5436	·0000116	+ 257	·0000002	+ 9	·0000000	0
0·4	·15542174	−146960	+4171	·0020284	+6744	·0000473	+ 554	·0000013	+ 37	·0000000	+ 2
0·5	·19146246	−175630	+4804	·0038574	+7666	·0001384	+ 973	·0000061	+ 98	·0000002	+ 10
0·6	·22574688	−199496	+5240	·0064530	+8160	·0003268	+1489	·0000207	+ 216	·0000014	+ 25
0·7	·25803635^{-}	−218121	+5449	·0098646	+8216	·0006641	+2046	·0000569	+ 408	·0000051	+ 67
0·8	·28814460	−231298	+5434	·0140978	+7848	·0012060	+2577	·0001339	+ 686	·0000155^{+}	+ 142
0·9	·31593987	−239040	+5215^{-}	·0191158	+7097	·0020056	+3002	·0002795^{-}	+1040	·0000401	+ 269
1·0	·34134475^{-}	−241568	+4815^{+}	·0248435^{+}	+6029	·0031054	+3259	·0005291	+1451	·0000916	+ 455
1·1	·36433394	−239280	+4272	·0311741	+4721	·0045311	+3283	·0009238	+1871	·0001886	+ 694
1·2	·38493033	−232721	+3625^{+}	·0379768	+3256	·0062851	+3051	·0015056	+2245	·0003550^{-}	+ 973
1·3	·40319952	−222536	+2914	·0451051	+1728	·0083442	+2560	·0023119	+2503	·0006187	+1243
1·4	·41924334	−209437	+2183	·0524062	+ 218	·0106593	+1834	·0033685^{-}	+2580	·0010067	+1453
1·5	·43319280	−194155^{-}	+1465^{-}	·0597291	−1201	·0131578	+ 931	·0046831	+2428	·0015400	+1525
1·6	·44520071	−177408	+ 793	·0669319	−2469	·0157494	− 82	·0062405^{+}	+2011	·0022258	+1399
1·7	·45543454	−159869	+ 191	·0738878	−3538	·0183328	−1114	·0079990	+1334	·0030515^{+}	+1012
1·8	·46406968	−142138	− 325^{-}	·0804899	−4386	·0208048	−2092	·0098909	+ 428	·0039784	+ 350
1·9	·47128344	−124733	− 744	·0866534	−4999	·0230676	−2930	·0118256	− 645	·0049403	− 574
2·0	·47724987	−108072	−1064	·0923170	−5381	·0250374	−3581	·0136958	−1798	·0058448	−1687
2·1	·48213558	− 92474	−1288	·0974425^{-}	−5549	·0266491	−3998	·0153862	−2934	·0065806	−2885
2·2	·48609655^{+}	− 78164	−1423	·1020131	−5528	·0278610	−4173	·0167832	−3950	·0070279	−4027
2·3	·48927589	− 65276	−1481	·1060309	−5348	·0286556	−4109	·0177852	−4761	·0070725^{-}	−4969
2·4	·49180246	− 53870	−1475^{+}	·1095139	−5045	·0290393	−3837	·0183111	−5296	·0066202	−5567
2·5	·49379033	− 43939	−1418	·1124924	−4651	·0290393	−3395	·0183074	−5517	·0056112	−5730
2·6	·49533881	− 35426	−1324	·1150058	−4201	·0286998	−2836	·0177520	−5414	·0040292	−5390
2·7	·49653303	− 28237	−1205^{-}	·1170991	−3720	·0280767	−2207	·0166552	−5008	·0019082	−4571
2·8	·49744487	− 22253	−1072	·1188204	−3237	·0272329	−1560	·0150576	−4345	−·0006699	−3331
2·9	·49813419	− 17340	− 934	·1202180	−2769	·0262331	− 943	·0130255^{-}	−3492	−·0035811	−1788
3·0	·49865010	− 13362	− 799	·1213387	−2331	·0251390	− 381	·0106442	−2520	−·0066711	− 97
3·1	·49903240	− 10183	− 671	·1222263	−1930	·0240068	+ 97	·0080109	−1511	−·0097708	+1589
3·2	·49931286	− 7675^{+}	− 555^{-}	·1229209	−1575	·0228843	+ 479	·0052265^{+}	− 528	−·0127116	+3105
3·3	·49951658	− 5722	− 451	·1234580	−1267	·0218097	+ 759	·0023893	+ 363	−·0153419	+4335
3·4	·49966307	− 4219	− 361	·1238684	−1005	·0208110	+ 947	−·0004116	+1123	−·0175387	+5188
3·5	·49976737	− 3078	− 285^{-}	·1241783	− 787	·0199070	+1049	−·0031002	+1724	−·0192167	+5622
3·6	·49984089	− 2221	− 221	·1244095^{+}	− 606	·0191079	+1079	−·0056164	+2156	−·0203325^{+}	+5643
3·7	·49989220	− 1586	− 170	·1245801	− 461	·0184167	+1053	−·0079170	+2423	−·0208840	+5304
3·8	·49992765^{+}	− 1120	− 128	·1247046	− 348	·0178308	+ 987	−·0099753	+2541	−·0209051	+4669
3·9	·49995190	− 783	− 96	·1247943	− 258	·0173436	+ 892	−·0117795^{+}	+2531	−·0204593	+3833
4·0	·49996833	− 541	− 70	·1248582	− 188	·0169456	+ 785	−·0133306	+2421	−·0196302	+2890
4·1	·49997934	− 370	− 52	·1249033	− 136	·0166261	+ 672	−·0146396	+2237	−·0185121	+1935
4·2	·49998665^{+}	− 251	− 37	·1249348	− 99	·0163738	+ 562	−·0157249	+2009	−·0172005^{-}	+1030
4·3	·49999146	− 168	− 26	·1249564	− 68	·0161777	+ 462	−·0166093	+1756	−·0157859	+ 239
4·4	·49999459	− 111	− 18	·1249712	− 48	·0160278	+ 368	−·0173181	+1498	−·0143474	− 407
4·5	·49999660	− 73	− 13	·1249812	− 34	·0159147	+ 292	−·0178771	+1249	−·0129496	− 893
4·6	·49999789	− 47	− 8	·1249878	− 22	·0158308	+ 224	−·0183112	+1023	−·0116411	−1223
4·7	·49999870	− 30	− 6	·1249922	− 16	·0157693	+ 172	−·0186430	+ 819	−·0104549	−1409
4·8	·49999921	− 19	− 4	·1249950^{+}	− 9	·0157250^{-}	+ 127	−·0188929	+ 645	−·0094096	−1473
4·9	·49999952	− 12	− 3	·1249969	− 7	·0156934	+ 94	−·0190783	+ 497	−·0085116	−1448
5·0	·49999971	− 8	+ 2	·1249981	− 5	·0156712	+ 69	−·0192140	+ 379	−·0077584	−1350
∞	·50000000	—	—	·1250000	—	·0156250	—	−·0195312	—	−·0051270	—

A table of numerical examples, designed to show the range of the new expansion, is given below. Various methods of checking were employed. Some are the examples of the earlier paper. Others, for lower values of N, were checked (a) by expansion in terms of multiple angles, or (b) from the manuscript of the *Tables of the Incomplete Beta-Function*. It would appear that seven-figure accuracy is obtainable by a computation which is easier than the previous one, and over a considerably larger range of N. The full expansion (4) was used for $N=9$ and $N=21$. In the case of $N=101$ the term in $\frac{1}{N^4}$ was dropped, while for the last example the term in $\frac{1}{N^3}$ was also neglected. The standard deviation (the unit being $\sin\theta$) is given.

N	σ	$\sin\theta$	x	$I_\theta(N)$	Correct result
9	·3015	·8	3	·499,1088	·499,1091
21	·2085	·4	1·912,8785	·473,6101	·473,6101
—	—	·729325	3·968,9483	·499,9737	·499,9737
101	·0985	·2	2·030,4889	·479,1034	·479,1033
—	--	·3	3·086,0353	·499,0127	·499,0127
401	·0498	·144	2·898,7034	·498,1373	·498,1373

It may be noted that the mathematician who desires not the probability integral, but merely $\int_0^\theta \cos^N\theta\, d\theta$, can readily compute it from the Table by the formula

$$\int_0^\theta \cos^N\theta\, d\theta=\sqrt{\frac{2\pi}{N}}\left\{\phi_0(x)-\frac{1}{N}\phi_1(x)+\frac{1}{N^2}\phi_2(x)-\frac{1}{N^3}\phi_3(x)+\frac{1}{N^4}\phi_4(x)\right\},$$

where $x=2\sqrt{N}\tan\frac{1}{2}\theta$.

Contributions to the Theory of Small Samples drawn from a finite Population*.

By J. SPLAWA-NEYMAN, Ph.D., Warsaw.

I. *Values of* β_1, β_2 *for the distribution of means of small samples taken from a finite population with any given frequency distribution.*

I understand by a *small sample* from a *finite* population one in which *all* the individuals in the sample are obtained at a single drawing, i.e. no individual is returned before another is drawn.

Let μ_k be the kth moment of the frequency curve of a population Ω, containing m individuals. The mean value of the character of the individuals considered we may suppose to be zero and the particular values we shall denote by

$$u_1,\ u_2,\ \dots\ u_m \quad\dots\dots(1),$$

so that

$$m\mu_1=\sum_{i=1}^{m} u_i=0 \quad\dots\dots(2).$$

Let as usual

$$\beta_1=\frac{\mu_3^2}{\mu_2^3} \quad\dots\dots(3);$$

$$\beta_2=\frac{\mu_4}{\mu_2^2} \quad\dots\dots(4).$$

---

* These results with others were originally published in *La Revue Mensuelle de Statistique*, publ. par l'Office Central de Statistique de la République Polonaise, tom. VI. pp. 1—29, 1923, but as that Journal may not be seen by English biometricians and as several important corrections have been made they are reproduced here.

We shall consider a sample ω_i of n individuals taken at random from the population Ω. It is clear that it is possible to have

$$N=\frac{m!}{n!\,(m-n)!}\quad\text{......}(5)$$

different samples ω_i, and that the i can be equal to 1, 2, ... N.

The values of the individual character considered in the sample ω_i we shall denote by

$$x_{i1},\ x_{i2},\ \dots\ x_{in}\quad\text{......}(6),$$

and their mean by x_i. Let us consider the population of N numbers x_i, where $i=1, 2, \dots N$, and calculate its four first moments about the mean, which we shall denote by M_k ($k=1, 2, 3, 4$). It is evident that

$$M_1=0\quad\text{......}(7).$$

If we take afterwards the ratios $B_1=\frac{M_3^2}{M_2^3},\quad B_2=\frac{M_4}{M_2^2}$ ......(8),

we shall be able to form some idea about the distribution of x_i if the distribution of (1) is given.

We shall begin with the calculation of M_2. Further, we shall denote by $E(y)$ the mean value of a variable y. Thus we have

$$\begin{aligned}M_2=E(x_i^2)&=\frac{1}{n^2}E\left(\sum_{k=1}^{n}x_{ik}^2+2\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}x_{ik}x_{il}\right)\\&=\frac{1}{n^2}\left(\sum_{k=1}^{n}E(x_{ik}^2)+2\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}E(x_{ik}x_{il})\right)\quad\text{......}(9).\end{aligned}$$

It is evident that $E(x_{ik}^2)=\mu_2$ ......(10),

$$E(x_{ik}x_{il})=\frac{2\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}u_ru_s}{m(m-1)}=-\frac{\sum_{r=1}^{m}u_r^2}{m(m-1)}=-\frac{\mu_2}{m-1}\quad\text{......}(11).$$

So, returning to (9), we have

$$M_2=\frac{1}{n^2}\left[n\mu_2-\frac{n(n-1)}{m-1}\mu_2\right]=\frac{m-n}{n(m-1)}\mu_2\quad\text{......}(12).$$

If we let m increase indefinitely, we shall have in the limit the well-known value for the second moment of the mean, namely

$$M_{2,\infty}=\frac{\mu_2}{n}\quad\text{......}(13).$$

Now we proceed to the third moment M_3. We have

$$\begin{aligned}M_3=E(x_i^3)&=\frac{1}{n^3}E\left[\sum_{k=1}^{n}x_{ik}^3+3\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(x_{ik}^2x_{il}+x_{il}^2x_{ik})+6\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{r=l+1}^{n}x_{ik}x_{il}x_{ir}\right]\\&=\frac{1}{n^3}\left[\sum_{k=1}^{n}E(x_{ik}^3)+3\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}E(x_{ik}^2x_{il}+x_{il}^2x_{ik})+6\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{r=l+1}^{n}E(x_{ik}x_{il}x_{ir})\right]\quad\text{......}(14).\end{aligned}$$

But

$$E(x_{ik}^3)=\mu_3\quad\text{......}(15),$$

$$E(x_{ik}^2x_{il})=\frac{\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}(u_r^2u_s+u_s^2u_r)}{m(m-1)}\quad\text{......}(16),$$

$$E(x_{ik}x_{il}x_{ir})=\frac{6\sum_{r=1}^{m-2}\sum_{s=r+1}^{m-1}\sum_{t=s+1}^{m}u_ru_su_t}{m(m-1)(m-2)}\quad\text{......}(17).$$

Evidently we have, using (3),

$$\left(\sum_{i=1}^{m}u_i\right)^3=\sum_{i=1}^{m}u_i^3+3\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}(u_r^2u_s+u_s^2u_r)+6\sum_{r=1}^{m-2}\sum_{s=r+1}^{m-1}\sum_{t=s+1}^{m}u_ru_su_t=0\quad\text{......}(18),$$

and

$$\sum_{i=1}^{m}u_i\sum_{i=1}^{m}u_i^2=\sum_{i=1}^{m}u_i^3+\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}(u_r^2u_s+u_s^2u_r)=0\quad\text{......}(19).$$

Hence
$$\sum_{r=1}^{m-1}\sum_{s=r+1}^{m}(u_r^2u_s+u_s^2u_r)=-m\mu_3 \qquad \text{.......(20)},$$

$$6\sum_{r=1}^{m-2}\sum_{s=r+1}^{m-1}\sum_{t=s+1}^{m}u_ru_su_t=2m\mu_3 \qquad \text{.......(21)}.$$

Putting these values in (16) and (17) and returning to (14), we have

$$M_3=\frac{1}{n^2}\left[\mu_3-\frac{3(n-1)}{m-1}\mu_2+\frac{2(n-1)(n-2)}{(m-1)(m-2)}\mu_3\right]=\frac{(m-n)(m-2n)}{n^2(m-1)(m-2)}\mu_3 \qquad \text{........}$$

a very symmetrical formula. It is evident that $M_3\rightarrow\frac{\mu_3}{n^2}$ if $m\rightarrow\infty$.

The calculation of M_4 follows in the same way. We have

$$\begin{aligned}M_4=E(x_i^4)=\frac{1}{n^4}E\Bigg[&\sum_{k=1}^{n}x_{ik}^4+4\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(x_{ik}^3x_{il}+x_{il}^3x_{ik})\\&+6\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}x_{ik}^2x_{il}^2+12\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{s=l+1}^{n}(x_{ik}^2x_{il}x_{is}+x_{il}^2x_{ik}x_{is}+x_{is}^2x_{ik}x_{il})\\&+24\sum_{k=1}^{n-3}\sum_{l=k+1}^{n-2}\sum_{s=l+1}^{n-1}\sum_{t=s+1}^{n}x_{ik}x_{il}x_{is}x_{it}\Bigg] \qquad \text{.......(23)}.\end{aligned}$$

Hence we have

$$E(x_{ik}^4)=\mu_4 \qquad \text{.......(24)},$$

$$E(x_{ik}^3x_{il})=\frac{\sum\limits_{r=1}^{m-1}\sum\limits_{s=r+1}(u_r^3u_s+u_s^3u_r)}{m(m-1)}=\frac{A}{m(m-1)}\ \text{(say)} \qquad \text{.......(25)},$$

$$E(x_{ik}^2x_{il}^2)=\frac{2\sum\limits_{r=1}^{m-1}\sum\limits_{s=r+1}^{m}(u_r^2u_s^2)}{m(m-1)}=\frac{B}{m(m-1)}\ \text{(say)} \qquad \text{.......(26)},$$

$$E(x_{ik}^2x_{il}x_{is})=\frac{2\sum\limits_{r=1}^{m-2}\sum\limits_{s=r+1}^{m-1}\sum\limits_{t=s+1}^{m}(u_r^2u_su_t+u_ru_s^2u_t+u_ru_su_t^2)}{m(m-1)(m-2)}=\frac{C}{m(m-1)(m-2)}\ \text{(say)} \ \text{...(27)},$$

$$E(x_{ik}x_{il}x_{is}x_{it})=\frac{24\sum\limits_{r=1}^{m-3}\sum\limits_{s=r+1}^{m-2}\sum\limits_{t=s+1}^{m-1}\sum\limits_{p=t+1}^{m}(u_ru_su_tu_p)}{m(m-1)(m-2)(m-3)}=\frac{D}{m(m-1)(m-2)(m-3)}\ \text{(say)} \ \text{...(28)}.$$

It is evident that

$$\left(\sum_{i=1}^{m}u_i\right)^4=m\mu_4+4A+3B+6C+D=0. \qquad \text{.(29)},$$

$$\sum_{i=1}u_i\sum_{i=1}u_i^3=m\mu_4+A=0 \qquad \text{.......(30)},$$

$$\left(\sum_{i=1}^{m}u_i\right)^2\sum_{i=1}^{m}u_i^2=m^2\mu_2^2+2A+C=0 \qquad \text{.......(31)},$$

$$\left(\sum_{i=1}^{m}u_i^2\right)^2=m\mu_4+B=m^2\mu_2^2 \qquad \text{.......(32)}.$$

Hence we have

$$A=-m\mu_4 \qquad \text{.......(33)},$$
$$B=m^2\mu_2^2-m\mu_4 \qquad \text{.......(34)},$$
$$C=2m\mu_4-m^2\mu_2^2 \qquad \text{.......(35)},$$
$$D=3m^2\mu_2^2-6m\mu_4 \qquad \text{.......(36)}.$$

Putting these values into (25), (26), (27) and (28) and returning to (23), we have

$$\begin{aligned}M_4&=\frac{1}{n^3}\Bigg[\mu_4-4\frac{n-1}{m-1}\mu_4+3\frac{n-1}{m-1}(m\mu_2^2-\mu_4)+\frac{6(n-1)(n-2)}{(m-1)(m-2)}(2\mu_4-m\mu_2^2)\\&\qquad+\frac{(n-1)(n-2)(n-3)}{(m-1)(m-2)(m-3)}(3m\mu_2^2-6\mu_4)\Bigg]\\&=\frac{m-n}{n^3(m-1)(m-2)(m-3)}[(m^2-6mn+m+6n^2)\mu_4+3m(m-n-1)(n-1)\mu_2^2] \qquad \text{......(37)}.\end{aligned}$$

We can now calculate B_1 and B_2. We have

$$B_1 = \frac{M_3^2}{M_2^3} = \frac{(m-1)(m-2n)^2}{n(m-2)^2(m-n)}\beta_1 \quad \text{......(38)},$$

$$\frac{M_4}{M_2^2} = \frac{(m-1)\left[(m^2-6mn+m+6n^2)\beta_2+3m(m-n-1)(n-1)\right]}{n(m-2)(m-3)(m-n)} \quad \text{......(39)}.$$

The formula (39) can be transformed into the following:

$$B_2 - 3\frac{m-1}{m+1} = \frac{m-1}{m(m-3)}\left[\frac{m+1}{m-2}\frac{(m-2n)^2}{n(m-n)} - 2\right]\left[\beta_2 - 3\frac{m-1}{m+1}\right] \quad \text{......(40)}.$$

If $m \to \infty$, so

$$B_1 \to B_{1\infty} = \frac{\beta_1}{n} \quad \text{......(41)},$$

$$B_2 \to B_{2\infty} = 3 + \frac{\beta_2 - 3}{n} \quad \text{......(42)}.$$

Let

$$k = 2\beta_2 - 3\beta_1 - 6 \quad \text{......(43)};$$

then, eliminating β_1 and β_2 from (41), (42) and (43), we have

$$2B_{2\infty} - 3B_{1\infty} - 6 = \frac{k}{n} \quad \text{......(44)}.$$

We can obtain another interesting result by eliminating n from (41) and (42). We have

$$B_{2\infty} = 3 + B_{1\infty}\frac{\beta_2 - 3}{\beta_1} \quad \text{......(45)},$$

a formula showing, that if we increase n, the point $M(B_{1\infty}, B_{2\infty})$ will approach the normal point (0, 3) following the right line from (β_1, β_2) to the normal point.

If we take

$$\frac{dB_{1\infty}}{dn} = -\frac{\beta_1}{n^2} \quad \text{......(46)},$$

$$\frac{dB_{2\infty}}{dn} = -\frac{\beta_2 - 3}{n^2} \quad \text{......(47)},$$

we see that the rates of decrease of $B_{1\infty}$ and $B_{2\infty}$ vary inversely with the square of n. We conclude that small values of n are sufficient to make these rates approach the normal point.

We also see from (43) and (44) that if in the classification of Pearson we distinguish only three types of curves depending upon the sign of the criterion k, the frequency distribution of the mean of the sample approaches the normal one without change of the type.

Mutatis mutandis these results can be extended to B_1 and B_2.

Certainly we find here some complications. We commence by eliminating

$$\frac{(m-2n)^2}{n(m-n)} \text{ from (38) and (40).}$$

Putting $\Delta = \beta_2 - 3\frac{m-1}{m+1}$, we have

$$B_2 - \left(3 - 2\frac{m^2(3+\Delta) - 9m - \Delta}{(m-3)m(m+1)}\right) = \frac{(m+1)(m-2)}{m(m-3)}\frac{B_1}{\beta_1} \quad \text{......(48)}$$

for the equation of the straight line which follows the point (B_1, B_2).

When n increases from 1 to $\frac{m}{2}$, or to $\frac{m-1}{2}$, this point starts from (β_1, β_2) and tends to the limit $M'\left(0, 3 - 2\frac{m^2(3+\Delta) - 9m - \Delta}{(m-3)m(m+1)}\right)$. Afterwards, if n still increases, the point (B_1, B_2) following the same way tends to (β_1, β_2). It is interesting to note that the formulae (38) and (40) do not change the values of B_1 and B_2 if we put $m-n$ instead of n. We see therefore that the distribution of means of samples containing n individuals has the same character as that of means of samples containing $m-n$ individuals.

Now consider the point $M'\left(0,\ 3-2\,\dfrac{m^2(3+\Delta)-9m-\Delta}{(m-3)\,m\,(m+1)}\right)$. We see that the ordinate differs from 3 only by a number of the order of $\dfrac{1}{m}$, and thus in most cases the point considered is practically the normal point. Its ordinate may be smaller or larger than 3 according to the sign of

$$\xi = m^2(3+\Delta)-9m-\Delta,$$

which may be positive or negative.

But this last case has no importance as ξ is negative only if

$$m<\frac{9+\sqrt{81+4\,(3+\Delta)\,\Delta}}{2\,(3+\Delta)}<10.$$

As the number of individuals in the general population is usually greater than 10, we can say that the ordinate of M' is as a rule less than 3. Now, if

$$k=2\beta_2-3\beta_1-6<0,$$

the straight line between the two points M' and $(\beta_1,\ \beta_2)$ does not cross the line

$$2B_2-3B_1-6=0,$$

and the type of the frequency distribution of means of samples is the same for all values of n. If $k \geqslant 0$, we have one, or two changes of type. The last case happens in the case when there is an integer value of $n>1$, for which

$$2B_2-3B_1-6=0,$$

where B_1 and B_2 are regarded as functions of n given by (38), (39).

Taking the differentials from (38) and (40), we have

$$\left.\begin{aligned}\frac{dB_1}{dn}&=-\frac{m^2(m-1)}{(m-2)^2}\,\frac{m-2n}{n^2\,(m-n)^2}\\ \frac{dB_2}{dn}&=-\frac{m^2-1}{(m-2)\,(m-3)}\,\frac{m\,(m-2n)}{n^2\,(m-n)^2}\left(\beta_2-3\,\frac{m-1}{m+1}\right)\end{aligned}\right\}\quad\text{.(49).}$$

We see that their absolute values decrease quickly to zero if n tends to $\tfrac{1}{2}m$, and we conclude that small values of n are sufficient to cause $(B_1,\ B_2)$ to approach the limit-point M'.

II. *The second moment of the squared standard deviation of a sample, the sample being taken at random from a finite population with a given distribution.*

The second moment required is given by the formula

$$\sigma_2{}^2=E(\sigma_t{}^4)-[E(\sigma_t)]^2.$$

Using the formulae (10) and (11), it is easy to calculate

$$E(\sigma_t{}^2)=E\left(\frac{1}{n}\sum_{k=1}^{n}x_{tk}{}^2-x_t{}^2\right)=\frac{n-1}{n}\left[E(x_{tk}{}^2)-E(x_{tk}x_{tl})\right]=\frac{(n-1)\,m}{n\,(m-1)}\,\mu_2\ \ldots\ldots\ldots(50).$$

Afterwards we observe that

$$\begin{aligned}\sigma_t{}^4&=\left(\frac{n-1}{n^2}\sum_{k=1}^{n}x_{tk}{}^2-\frac{2}{n^2}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}x_{tk}x_{tl}\right)^2\\ &=\frac{(n-1)^2}{n^4}\left(\sum_{k=1}^{n}x_{tk}{}^4+2\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}x_{tk}{}^2x_{tl}{}^2\right)\\ &\quad-4\,\frac{n-1}{n^4}\left[\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(x_{tk}{}^3x_{tl}+x_{tk}x_{tl}{}^3)+\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{s=l+1}^{n}(x_{tk}{}^2x_{tl}x_{ts}+x_{tk}x_{tl}{}^2x_{ts}+x_{tk}x_{tl}x_{ts}{}^2)\right]\\ &\quad+\frac{4}{n^4}\left[\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}x_{tk}{}^2x_{tl}{}^2+2\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{s=l+1}^{n}(x_{tk}{}^2x_{tl}x_{ts}+x_{tk}x_{tl}{}^2x_{ts}+x_{tk}x_{tl}x_{ts}{}^2)\right.\\ &\qquad\left.+6\sum_{k=1}^{n-3}\sum_{l=k+1}^{n-2}\sum_{s=l+1}^{n-1}\sum_{t=s+1}^{n}x_{tk}x_{tl}x_{ts}x_{tt}\right]\end{aligned}$$

$$=\frac{(n-1)^2}{n^4}\sum_{k=1}^{n}x_{ik}^4+2\frac{n^2-2n+3}{n^4}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}x_{ik}^2x_{il}^2-4\frac{n-1}{n^4}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(x_{ik}^3x_{il}+x_{ik}x_{il}^3)$$

$$-4\frac{n-3}{n^4}\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{s=l+1}^{n}(x_{ik}^2x_{il}x_{is}+x_{ik}x_{il}^2x_{is}+x_{ik}x_{il}x_{is}^2)$$

$$+\frac{24}{n^4}\sum_{k=1}^{n-3}\sum_{l=k+1}^{n-2}\sum_{s=l+1}^{n-1}\sum_{t=s+1}x_{ik}x_{il}x_{is}x_{it}.$$

Using the formulae (24)—(36) we have

$$E(\sigma_i^4)=\frac{1}{n^3}\left[(n-1)^2\mu_4+\frac{(n^2-2n+3)(n-1)}{m-1}(m\mu_2^2-\mu_4)+4\frac{(n-1)^2}{m-1}\mu_4\right.$$

$$\left.-2\frac{(n-1)(n-2)(n-3)}{(m-1)(m-2)}(2\mu_4-m\mu_2^2)+3\frac{(n-1)(n-2)(n-3)}{(m-1)(m-2)(m-3)}(m\mu_2^2-2\mu_4)\right]$$

$$=\frac{m(m-n)(mn-m-n-1)(n-1)}{n^3(m-1)(m-2)(m-3)}\mu_4-\frac{m(n-1)[m^2(n^2-2n+3)-m(3n^2+3)+3n^2+3n]}{n^3(m-1)(m-2)(m-3)}\mu_2^2 \qquad \text{..................................(51).}$$

If we square (50) and subtract it from (51), we have

$$\sigma_2^2=\frac{m(m-n)(n-1)}{n^3(m-1)^2(m-2)(m-3)}\mu_2^2[(mn-m-n-1)(m-1)\beta_2-(m^2n-3m^2+6m-3n-3)]\text{...(52),}$$

a formula giving the squared standard deviation of the squared standard deviations in samples.

This result, being a generalisation of formulae given by other authors, is, I believe, novel and of considerable importance. It agrees for m indefinitely large with the value given by "Student" in *Biometrika*, Vol. VI. p. 3. For he has

$$M_1'=\mu_2\frac{n-1}{n},\qquad M_2'=\mu_4\frac{(n-1)^2}{n^3}-\mu_2^2\frac{(n-1)(n^2-2n+3)}{n^3}$$

and therefore

$$\sigma_2^2=\mu_4\frac{(n-1)^2}{n^3}-\mu_2^2\frac{(n-1)(n-3)}{n^3}.$$

See also Tchouproff, *Biometrika*, Vol. XII. p. 193, and Church, *ibid.* Vol. XVII. p. 82.

III. *The correlation between the square of the deviation of the mean of a sample from the mean of the sampled population with the square of the standard deviation, the sample being taken at random from a finite population with any given distribution.*

Let R be the required coefficient of correlation. We shall have

$$R=\frac{E(x_i^2\sigma_i^2)-E(x_i^2)E(\sigma_i^2)}{\sqrt{\{E(x_i^4)-[E(x_i^2)]^2\}\{E(\sigma_i^4)-[E(\sigma_i^2)]^2\}}} \qquad \text{........................(53).}$$

It is evident that

$$x_i^2\sigma_i^2=x_i^2\left(\frac{1}{n}\sum_{k=1}^{n}x_{ik}^2-x_i^2\right)=\frac{1}{n^3}\left(\sum_{k=1}^{n}x_{ik}\right)^2\sum_{k=1}^{n}x_{ik}^2-x_i^4$$

$$=\frac{1}{n^3}\left[\sum_{k=1}^{n}x_{ik}^4+2\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(x_{ik}^2x_{il}^2+x_{ik}^3x_{il}+x_{ik}x_{il}^3)\right.$$

$$\left.+2\sum_{k=1}^{n-2}\sum_{l=k+1}^{n-1}\sum_{s=l+1}^{n}(x_{ik}^2x_{il}x_{is}+x_{ik}x_{il}^2x_{is}+x_{ik}x_{il}x_{is}^2)\right]-x_i^4 \quad \text{...(54).}$$

Hence, to calculate $E(x_i^2\sigma_i^2)$, we may use the calculated means of x_{ik}^4, $x_{ik}^2x_{il}^2$, $x_{ik}^3x_{il}$ and $x_{ik}^2x_{il}x_{is}$, namely (24)—(27) and (37). We shall have

$$E(x_i^2\sigma_i^2)=\frac{1}{n^3}\left[\mu_4+\frac{n-1}{m-1}(m\mu_2^2-\mu_4)-2\frac{n-1}{m-1}\mu_4+\frac{(n-1)(n-2)}{(m-1)(m-2)}(2\mu_4-m\mu_2^2)\right]-M_4$$

$$=\left[\frac{(m-n)(m-2n)}{n^3(m-1)(m-2)}-\frac{(m-n)(m^2-6mn+m+6n^2)}{n^3(m-1)(m-2)(m-3)}\right]\mu_4$$

$$+\left[\frac{(m-n)(n-1)}{n^3(m-1)(m-2)}-3\frac{(m-n)(m-n-1)(n-1)}{n^3(m-1)(m-2)(m-3)}\right]m\mu_2^2$$

$$=\frac{m(m-n)(m-2n+1)(n-1)}{n^3(m-1)(m-2)(m-3)}\mu_4+\frac{m(m-n)(mn-3m+3)(n-1)}{n^3(m-1)(m-2)(m-3)}\mu_2^2 \quad \text{.........(55)}$$

Now from (12)
$$E(x_i^2)=M_2=\frac{m-n}{n(m-1)}\mu_2 \quad \text{.................................(56)},$$

and it is easy to calculate

$$E(\sigma_i^2)=E\left(\frac{1}{n}\sum_{k=1}^{n}x_{ik}^2-x_i^2\right)=\frac{n-1}{n}\left[E(x_{ik}^2)-E(x_{ik}x_{il})\right]=\frac{(n-1)m}{n(m-1)}\mu_2 \quad \text{.........(57)},$$

using the same formulae (10) and (11). Multiplying (56) and (57) and subtracting from (55), we have

$$\begin{aligned}R\sigma_1\sigma_2&=\frac{m(m-n)(m-2n+1)(n-1)}{n^3(m-1)(m-2)(m-3)}\mu_4-\frac{m(m-n)[3m^2-m(4n+6)+6n+3](n-1)}{n^3(m-1)^2(m-2)(m-3)}\mu_2^2\\ &=\frac{m(m-n)(n-1)}{n^3(m-1)^2(m-2)(m-3)}\mu_2^2\{(m-2n+1)(m-1)\beta_2-[3m^2-m(4n+6)+6n+3]\} \quad \text{...(58)},\end{aligned}$$

where σ_1, σ_2 are the standard deviations of x_i^2 and σ_i^2 respectively. We have further

$$\begin{aligned}\sigma_1^2&=E(x_i^4)-[E(x_i^2)]^2=M_4-M_2^2\\ &=\frac{m-n}{n^3(m-1)(m-2)(m-3)}[(m^2-6mn+m+6n^2)\mu_4+3m(m-n-1)(n-1)\mu_2^2]-\frac{(m-n)^2}{n^2(m-1)^2}\mu_2^2\\ &=\frac{(m-n)\mu_2^2}{n^3(m-1)^2(m-2)(m-3)}[(m^2-6mn+m+6n^2)(m-1)\beta_2+2n(m-n)(m^2+m-3)\\ &\qquad -3m(m-1)^2] \quad \text{......(59)}.\end{aligned}$$

Now, putting the values (58), (59) and (52) in the equation (53), we have

$$R=\frac{\sqrt{m(n-1)}\,[(m-2n+1)(m-1)\beta_2-\{3m^2-2m(2n+3)+3(2n+1)\}]}{\sqrt{\left\{\begin{array}{r}[(m^2-6mn+m+6n^2)(m-1)\beta_2+2n(m-n)(m^2+m-3)-3m(m-1)^2]\\ \times[(mn-m-n-1)(m-1)\beta_2-(m^2n-3m^2+6m-3n-3)]\end{array}\right\}}}$$
$$\text{.......................(60)}.$$

As the above formula is very complicated it is desirable to give some approximations. The first we reach by assuming that m is indefinitely large, but n finite. Dividing the numerator and the denominator of R by $m^{\frac{5}{2}}$ and increasing m indefinitely, we have

$$R'=\frac{\sqrt{n-1}\,(\beta_2-3)}{\sqrt{(\beta_2+2n-3)[(n-1)\beta_2-n+3]}} \quad \text{............................(61)}.$$

But it is also possible to find an approximation for a finite sample from a finite population. I followed at first the usual method, expanding R in a series in terms of $1/m$, and retaining only the two first terms, but this method has two disadvantages. First the formula reached is as complicated as (60), and secondly it assumes that not only a/m^2 is negligible, where a is of the order of a single digit, but also that all ratios n^p/m^2 can be neglected. It seemed better therefore to suppose all ratios $9/m$ or less and also β_2/m were negligible and retain n/m and its powers. We thus reach a very simple approximative formula:

$$R''=-\frac{\dfrac{1}{m}-\dfrac{\beta_2-3}{2n}\left(1-\dfrac{2n}{m}\right)}{\sqrt{\left(1-\dfrac{n}{m}+\dfrac{\beta_2-3}{2n}\right)\left(\dfrac{1}{n-1}+\dfrac{\beta_2-3}{2n}\right)}} \quad \text{.......................(62)}.$$

If $\beta_2=3$, we have

$$R''=-\sqrt{\frac{n}{m(m-n)}} \quad \text{...(63)}.$$

IV. ***The correlation between the mean of a sample and its squared standard deviation, the sample being taken at random from a finite population.***

In the previous notation x_i was the deviation of the mean of the ith sample from the mean of the sampled population. Thus the correlation required is given by

$$\rho=E[x_i(\sigma_i^2-E(\sigma_i^2))]/\sigma_2\sqrt{M_2} \quad\text{..................................(64).}$$

But as

$$E[x_i E(\sigma_i^2)]=E(x_i)\,E(\sigma_i^2)=0,$$

we have

$$\rho=E[x_i\sigma_i^2]/\sigma_2\sqrt{M_2} \quad\text{..................................(65).}$$

As σ_2 and M_2 have been calculated already, we need only find $E[x_i\sigma_i^2]$. Now,

$$x_i\sigma_i^2=\frac{1}{n^2}\left(\sum_{k=1}^{n} x_{ik}\right)\left(\sum_{k=1}^{n} x_{ik}^2\right)-x_i^3$$

$$=\frac{1}{n^2}\left\{\sum_{k=1}^{n} x_{ik}^3+\sum_{k=1}^{n-1}\sum_{l=k+1}^{n}(x_{ik}^2x_{il}+x_{ik}x_{il}^2)\right\}-x_i^3,$$

and therefore

$$E[x_i\sigma_i^2]=\frac{1}{n}E(x_{ik}^3)+\frac{n-1}{n}E(x_{ik}^2x_{il})-M_3,$$

whence using the formulae (15), (16), (20), and (22), we find easily

$$E[x_i\sigma_i^2]=\frac{m(m-n)(n-1)}{n^2(m-1)(m-2)}M_3 \quad\text{..................................(66).}$$

Putting this result into (65) and using (12) and (52), we have finally

$$\rho=\frac{\sqrt{m(m-1)(m-3)(n-1)}\,\sqrt{\beta_1}}{\sqrt{(m-2)\{(mn-m-n-1)(m-1)\beta_2-(m^2n-3m^2+6m-3n-3)}} \quad\text{.........(67).}$$

This result is a generalisation, now first published, of one previously reached. If m be made indefinitely large, we have in the limit,

$$\rho'=\frac{\sqrt{n-1}\,\sqrt{\beta_1}}{\sqrt{(n-1)\beta_2-n+3}} \quad\text{..................................(68),}$$

and for a large sample*,

$$\rho''=\sqrt{\beta_1}/\sqrt{\beta_2-1} \quad\text{..................................(69).}$$

We see that, if the frequency distribution of the original population be symmetrical and thus $\beta_1=0$, the coefficient of correlation between the mean and the squared standard deviation is always zero. It would be erroneous, however, to assume that they are independent. The only known case of independency of the mean and the standard deviation is when the original population is normal and indefinitely large. The above results enable us to prove that if the original population be indefinitely large, but not normal, the standard deviation and the mean of a sample, and also all their powers, are not independent. This result seems to me important, because it shows that the normal curve is the only curve by which, knowing the frequency distribution $y=f(x)$ of the mean of a sample and the frequency distribution $y=\psi(\sigma)$ of the standard deviation, we reach the frequency surface of means and standard deviations simply by multiplying

$$F(x,\sigma)=f(x)\times\psi(\sigma).$$

In all other cases such multiplying would be illegitimate.

---

* Cf. *Biometrika*, Vol. IX. p. 7, Equation (xxvi).

CAMBRIDGE: PRINTED BY
W. LEWIS, M.A.
AT THE UNIVERSITY PRESS

I. A. R. I. 75.

IMPERIAL AGRICULTURAL RESEARCH
INSTITUTE LIBRARY
NEW DELHI.

Date of issue.	Date of issue.	Date of issue.
11-9-41		
22-9-41		
15.11.43		
13.10.61		
10/8/63		